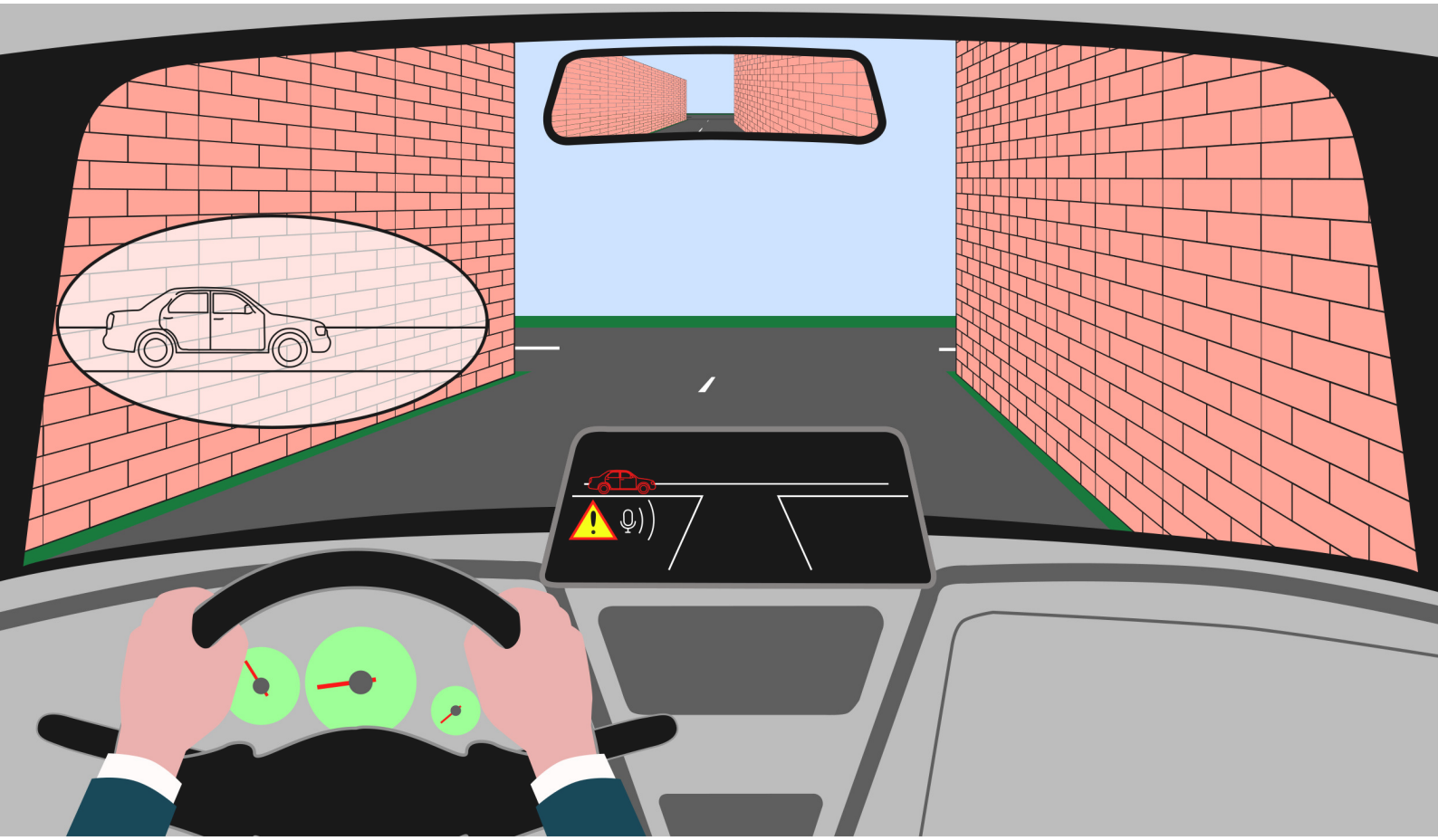
TU Delft

Acoustic Non-Line-of-Sight Vehicle Localization in Urban Environments



by

Yannick Schulz

to obtain the degree of Master of Science
at the Delft University of Technology,

Student number: 4757920
Date: September 7, 2021
Supervisors: Dr. J. F. P. Kooij, TU Delft
T. M. Hehn, TU Delft

Preface

The new decade has been dominated so far by one major news that riddled the entire world. The SARS-CoV-2 pandemic brought suffering to many people and their families. It disrupted the economy, social life and didn't stop at the academic wing. Many schools and universities struggled to provide the education and opportunities they wished to. I will always connect the work on this thesis with these challenging times, but also with the great work of all the people part of the scientific community that worked day and night to pave the way towards a better future and I am beyond grateful for that.

The idea of using acoustic localization techniques in driving applications stems from my supervisor Dr. Julian F.P. Kooij. Already in his time at the University of Amsterdam, when he was working on merging audio-visual features in the context of pedestrian surveillance, he had the vision to bring this concept to mobile vehicle platforms. With his position as Assistant Professor of the Cognitive Robotics (CoR) research group at the TU Delft, he was able to provide this opportunity to me as his Master's student to endeavour on this journey. As a novel modality to vehicular sensing, I had to accept to take part in this exciting challenge and I want to express my gratitude to him for this. Furthermore, I want to thank Avinash K. Mattar, a fellow Masters student, who started a parallel topic alongside mine using acoustic sensing, and our second supervisor Thomas M. Hehn in particular. Together we have spent hours gathering and processing data and working towards a publication in the renowned Robotics and Automation Letters (RA-L). Immaculate teamwork and an inspiring dynamic made the days in blazing heat, driving up and down an avenue and nightly zoom calls not only bearable but satisfying. As the accompanying paper has already been published, many thanks also go out to the organizers, associates and reviewers of the RA-L Journal and International Conference on Robotics and Automation (ICRA) for challenging and accepting our submission.

Last but not least I want to thank and remember Prof. Dr. Hans Müller-Storz. He was instrumental in my inspiration to pursue a technical career many years ago. Before he passed away, he was leading the biomechanical laboratory for vibration theory at the Hochschule Offenburg. There he introduced me to bicycle dynamics in the scope of a high school project when I was a mere pupil.

Yannick Schulz
Delft, September, 2021

Contents

Symbols

Symbols

$\mathscr{C}$	Set of all classes
x, X	Audio signal in time and frequency domain
f, ω	Frequency and angular frequency
t, τ	Time and discrete time variable
m, n	Arbitrary microphones or pair (m, n)
$C, c_{\_}$	Cross correlation matrix and element
Ψ	Filter function
ϕ	Azimuth angle
Δ	Steering delay
c	Speed of sound, $c = 343\mathrm{m/s}$
$S, s_{\_}$	Steered correlation matrix and element
M	Number of all microphones
$\hat{M}$	Number of all microphone pairs
F	Number of all frequency bins
T	Number of all time bins
L	Number of segments
B	Number of angular bins
$\mathbf{u}$	Feature vector
y	Class label
λ	ℓ_2 regularization parameter
$\mathbf{w}, w_{\_}$	Classifier parameter vector or element
μ	Mean
σ	Variance
v	Video frame number
t_0	Alignment and label time for static recordings
$\widetilde{t}_0$	Alignment and label time for dynamic recordings
N	Number of samples
δt	Sample length
J_3	Criterion based on scatter matrices
Σ	Scatter matrix

Acronyms and Abbreviations

CoR	Cognitive Robotics
RA-L	Robotics and Automation Letters
ICRA	International Conference on Robotics and Automation
NLOS	Non Line of Sight
LOS	Line of Sight
LiDAR	Light Detection and Ranging
RADAR	Radio Detection and Ranging
SONAR	Sound Navigation and Ranging

SSL	Sound Source Localization
DAS	Delay and Sum
DOA	Direction of Arrival
TDOA	Time Difference of Arrival
MVDR	Minimum Variance Distortionless Response
MUSIC	Multiple Signal Classification
ESPRIT	Estimation of Signal Parameter Invariance Technique
TOPS	Test of Projected Subspaces
CSSM	Coherent Signal-Subspace Processing
ML	Maximum Likelihood
LOCATA	Localization and Tracking
DCASE	Detection and Classification of Acoustic Scenes and Events
CNN	Convolutional Neural Network
SELD	Sound Event Localization and Detection
MAV	Micro Aerial Vehicle
SNR	Signal to Noise Ratio
SLAM	Simultaneous Localization and Mapping
ITD	Inter-aural Time Difference
ILD	Inter-aural Level Difference
SRP-PHAT	Steered Response Power Phase Transform
GCC-PHAT	Generalized Cross Correlation Phase Transform
STFT	Short Time Fourier Transform
UW-SVM	Uniformly Weighted Support Vector Machine
SVM	Support Vector Machine
SW-CNN	Selectively Weighted Convolutional Neural Network
DC	Direct Current
MEMS	Microelectromechanical System
ROS	Robot Operating System
FOV	Field of View
t-SNE	t-Distributed Stochastic Neighbor Embedding
COCO	Common Objects in Context
GPS	Global Positioning System

Abstract

Driving is a challenging task. When people operate vehicles they utilize all their senses to assess the current traffic scenario and determine appropriate actions to take. Sensors in autonomous driving applications aim to mimic those human senses to build a similar understanding of these complex circumstances. Most scientific attention in the autonomous driving community has been put on camera and LiDAR solutions, which are already capable of constructing a cohesive three-dimensional representation of the surroundings with high accuracy. Instead, minimal research went into the utilization of the auditory landscape in traffic scenarios. With the use of sirens in emergency vehicles and upcoming regulations of minimum sound emissions for electric vehicles, it appears evident that acoustic perception deserves more attention. Auditory localization in particular is a task widely studied in the field of speaker recognition, where robust methods exist that can localize and track multiple speech sources in difficult acoustic environments. This raises the question of whether such principles can be applied in the domain of autonomous driving and help to make self-driving cars safer and more robust in navigating dense and complex urban cities. Especially in non-line-of-sight situations where oncoming traffic can be hazardous and conventional systems fail, employing acoustic sensing may prove instrumental in early detection and localization attempts. This study investigates, if, and to what extend, acoustic methods can be used to complement ordinary sensing and localization methods. Using the example of a T-crossing, where traffic is occluded by buildings, it is shown that acoustic sensing alone is capable of detecting oncoming traffic before it enters direct line-of-sight. To achieve this, a combination of a generic acoustic line-of-sight feature and two concepts of data-driven classification methods are used to infer from the surrounding soundfield at the ego-vehicle to other motorized vehicles in proximity. This investigation is aided by a supplementary real-world dataset that provides around two hours of data, gathered from five different locations. The performance of the methods is directly compared to a visual baseline and other acoustic line-of-sight methods. The results demonstrate that the proposed acoustic localization concepts can detect traffic about one second earlier than conventional line-of-sight sensors. Despite the complexity of the problem, it is shown that the more lightweight method in terms of parameter count is favourable and more performant in most of the tasks. Among others, these tasks include generalization across different environments and a larger time horizon.

1 Introduction

The problem of autonomous driving in urban environments, particularly small street canyons in European historic city centers, is significantly greater than in other traffic scenarios including interstate road networks such as motorways [1]. Cities that have developed over time and existed prior to the golden age of motorized passenger cars present a unique challenge for today's infrastructure enhancement. Limited by the historic street layout, smart solutions are required to accommodate the variety of today's transportation spectrum. Due to the historic street layout, innovative solutions are necessary to handle the variety of today's transportation spectrum. As a result, cities are densely populated with a variety of road users, such as pedestrians, cyclists, scooters, trams, and many more. Apart from residential and commercial demands, all of these forms of transportation need their own infrastructure and real estate inside a modern city, narrowing the usable road width for passenger cars. As a result, the environment becomes extremely cluttered, making it difficult for both humans and machines to navigate.

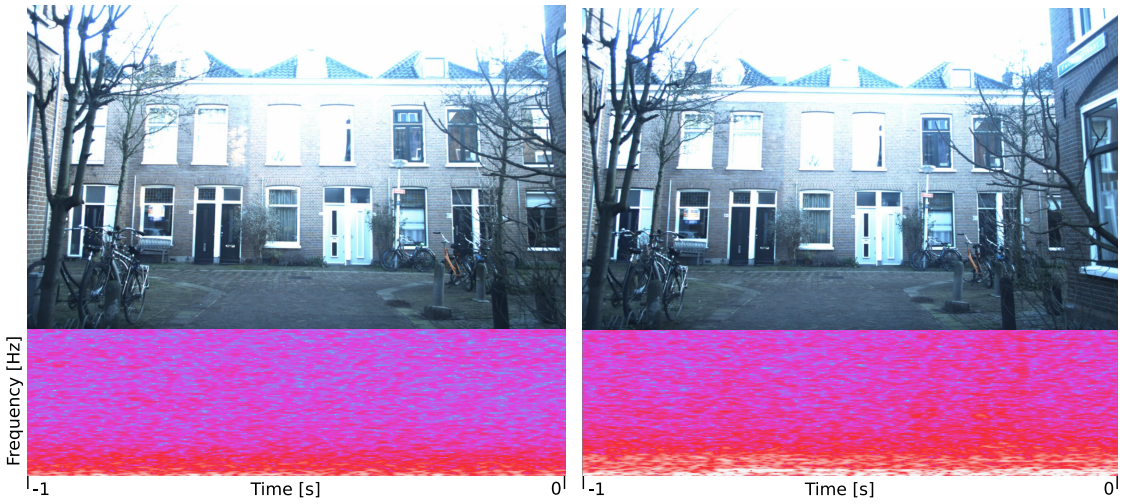
1.1. Problem Statement

Current vehicle platforms that operate in a higher level of automation, commonly use an array of different sensor applications. Most notable are camera or stereo camera pairs, LiDAR and RADAR sensors [2]. Each of which with its own set of pros and cons, they still have one thing in common. They detect a signal from a source or a reflection in their sensory field of view (FOV). A direct line-of-sight (LOS) towards the target is required to confidently correlate a signal to its source. As a result, when the target is partially obscured or beyond the sensor's range of vision, it performs poorly. While many studies have addressed these issues and iterate in improving the inference with incomplete data, for example in segmentation algorithms [3] [4], visual obstructions are undoubtedly malicious for any kind of method using any of these sensors. Important events or road users might be easily overlooked, whereas early detection is a pillar to avoid hazardous situations. Comparing these state of the art perception systems with human perception, one has benefits the other one has not. Humans are unique in their ability to compensate for and extrapolate from limited input and can rely on other senses if others are impaired due to occlusions. Humans, in particular, may utilize spatial hearing to pinpoint certain events or targets that generate noise with or without a visual line-of-sight [5]. This modality has experienced little attention in the autonomous driving community. More precisely, the use of the urban soundscape does not feature in this rapidly developing branch to add more fidelity in navigation or increase safety. However, vehicle sounds are a subconscious indicator for a lot of drivers as the soundfield of a traffic scenario exhibits undoubtedly a lot of information [6]. Being able to localize approaching vehicles early could greatly improve the robustness of navigating highly obstructed terrain. Following this line of reasoning, it can be stated that acoustic cues can also be used in automated systems, localizing sound sources in their vicinity.

Although sound source localization (SSL) has not been fully explored yet in this context, other applications have thrived employing methods to localize events in certain environments. Speaker localization [7] in meeting rooms or road and air traffic surveillance are popular examples. Acoustic localization can also be used under water. In ocean acoustics whales or vessels can be localized and tracked over large distances [8]. Because of their environment, these applications are well suited for sound source localization methods. Firstly, a stationary sensor removes the degree of freedom of a changing acoustic scene that can vary

heavily in dynamic scenarios. But secondly, they show how sound source localization can be used over large distances due to the reflection and diffraction properties of soundwaves. As soundwaves travel through a medium they get easily reflected at medium borders, amplifying the magnitude of the wavefront due to interference. Soundwaves travel particularly far in the ocean in this manner, reverberating on the surface and ground. Lower frequency soundwaves can bend around narrow objects or pass around wedges of large buildings due to diffraction in road and air traffic monitoring. Reflection and diffraction of soundwaves also occur on a much smaller scale like in traffic, which makes it possible to hear oncoming vehicles without visual sight. On a much smaller scale, such as in traffic, sound reflection and diffraction occur as well, allowing drivers to hear oncoming cars without visual sight.

Applying this concept in the context of autonomous driving applications, one can assume the following: In these situations of visually occluded traffic events or road users that emit sound, the soundfield can reach an ego vehicle that lies in the shadow zone of the occlusion. Figure 1.1 visualizes this concept, along with the spectrogram of a one-second audio sample captured beside it. Displayed is a generic traffic scene including the spectrogram of a one-second audio sample that was recorded alongside. No other road users are in the vicinity of the sensors in Figure 1.1a. However, in Figure 1.1b a car is approaching from around the left corner. While the camera picture remains the same, the spectrogram detects a signal of lower frequencies and a change of the soundfield.



(a) No traffic is in proximity

(b) Traffic is coming from around the left corner

Figure 1.1: A snapshot of an arbitrary traffic scene with the audio spectrogram (0 to 8000Hz) of a one second sample is displayed. In 1.1a no vehicle is present, while in 1.1b a vehicle is approaching from behind the left corner.

The primary difference between the aforementioned and established approaches is that without a straight line-of-sight, the soundfield's directivity is poorly defined. In non-line-of-sight (NLOS) conditions the soundfield is merely a superposition of reflected and diffracted components of the initial soundwave and susceptible to the scene's complicated acoustic structure. Although knowing the dominant direction of the soundfield at the given point of measurement only determines the direction of the reflected or diffracted wave which may not point to the actual sound source. Multiple microphones must be used to record the same signal at different points in order to establish the direction of a soundfield. Although binaural systems are available, microphone arrays typically consist of four or more microphones. However, using more than two microphones is recommended. Not only does it increase the number of observable directions (depending on the distribution of the microphones), but it also makes the array more robust against internal and external disturbances. The time difference of arrival (TDOA) of the recorded signal correlates to the direction of the current soundfield.

In conclusion, the main challenge of this thesis is to deploy a microphone array to localize oncoming traffic in these adverse conditions of an occluded target vehicle. An appropriate technique has to be developed that can utilize the reverberated soundfield and extract information about the sound source location.

1.2. Research Questions

To elaborate, small street canyons leading to junctions that are difficult to analyze due to occlusions are investigated in the scope of this thesis. Especially when approaching a T-crossing (as the one displayed in Figure 1.1), the traffic situation and hence oncoming vehicles are impossible to detect for conventional sensors until the very last meters. But oncoming vehicles from the left and right of the continuous section can quickly pose a hazard when manoeuvring these walled-off environments. The advantages that could be achieved in this scenario and the possibility of success are compelling for further investigation. Therefore, it is proposed to use a purely acoustical method aided by a vehicle-mounted microphone array (see Figure 1.2) to explore the localization capabilities of this concept in real world environments. The proposed method should be able to detect and localize oncoming road vehicles before they enter line-of-sight. It should also be able to differentiate if a car is in proximity or not and, if so, from which direction it is approaching. Due to the complexity and irregularities in real world conditions shaping the soundfield, it is proposed to use learning based techniques above the conventional, purely signal processing, techniques. Signal Processing techniques have been historically used as go-to methods for LOS acoustic localization applications. A comparison between those and the proposed methods shall reveal the benefits of learning based techniques in these complex environments. Despite concentrating on the overall architecture of a T-crossing, not all T-crossings may be regarded acoustically identical. Changes in road widths, wall surface texture, and the position of the ego vehicle can all have a major impact on the acoustic response. Based on these assumptions and conclusions the following research questions are designed:

- How accurately can a purely acoustic localization system predict oncoming traffic in the specified scenario?
- How do data-driven methods compare to classical signal processing methods for acoustic localization in non-line-of-sight situations?
- How does a different location or ego-motion of the sensor array influence the accuracy of the approach?
- How much earlier can the proposed method confidently detect traffic with respect to a visual method?
- Is an end-to-end approach viable without any preprocessing of the data for the task at hand?
- Are there limits to the proposed methods and what are potential areas for improvements?

Figure 1.2: Display of the research vehicle with the mounted microphone array on the roof.

The thesis is structured as follows: First, a short survey of existing SSL techniques is reported in Chapter 2. Both, signal processing and data-driven techniques are shown and their strengths and weaknesses are discussed. The chapter is concluded by a summary of the novel contributions this thesis puts forward. Chapter 3 introduces the methods that are used to solve the problems depicted here. Starting with a more specific problem definition, a detailed overview of the exact algorithms are provided. The second half of the Methodology chapter is dedicated towards the data acquisition that facilitates this research. The following Chapter 4 includes the experiments that provide the basis of reasoning. First, qualitative results are shown, together with a definition of the metrics that are used to judge the performance of the methods. A section follows that validates the choices that were made considering the specifics of the methods. The rest of the chapter is comprised of the experiments that are used to answer the aforementioned research questions. Lastly, Chapter 5 discusses the results of the experiments and puts them into perspective. The research questions are addressed and answered. This chapter is concluded by a perspective for future work that addresses improvements of the proposed methods and beyond. Appendix A includes the accompanying research paper that was published in the Robotics and Automation Letters.

2

Related Work

The transition to a higher ratio of low-noise electrically powered cars on public roadways, as part of the journey to a more sustainable future, raises a fundamental question about practicality that must be addressed immediately. Traffic noise pollution has long been a source of annoyance and concern for public health. As a result, this subject has received scientific attention for as long as motorized vehicles have existed. A common misconception about traffic noise is about the origin thereof. Studies show that three types of sound generation are the core of what we perceive as traffic noise. These are engine, tyre and aerodynamic noise [9]. The misconception therein is, that the engine noise is the most dominant source. While this is true for a range of vehicle types, such as lorries, motorcycles, and sports cars, this is not true for the majority of the fleet that makes up all motorized road vehicles. In reality, engine noise is only dominant at very low vehicle speeds, while tyre noise supersedes at urban driving speeds of 30km/h to 50km/h. Well above these speeds aerodynamic noise tends to overcome tyre noise ultimately, but at such high speeds that it becomes irrelevant for the problem at hand.

2.1. Acoustic Localization Techniques

Acoustic localization is not a new concept by any means. Sound navigation and ranging (SONAR) has a long history in naval navigation. A transponder sends out sound waves, while a detector captures echoes of these signals, creating a spatio-temporal picture of the surrounding objects. Similar to an active ranging system, Sound Source Localization is the passive equivalent. The target itself emits sound signals while a sensor, commonly a detector array, is used to infer to the direction it is coming from.

2.1.1. Model Based Techniques

Signal processing techniques have emerged in various fields as this setup can not only apply to microphone, but also antenna or other arrays. Most commonly known are beamforming techniques, such as the delay-and-sum (DAS) beamformer [10]. It is a spatial filtering technique to condense the direction of arrival (DOA) of a signal source. A set of candidate directions are proposed around the array. For each so-called steering angle, the M microphone signals get reversely delayed according to the time difference of arrival. All signals are delayed to each other, corresponding to a microphone pair, and its TDOAs and subsequently summed. The superposition of the delayed signals is only then constructive if the direction coincides with the actual direction the signal is coming from. The result is a mesh of DOA magnitudes for each sampled direction, depending on the sampling arrangement of directions. This beamforming technique has been adapted and adjusted throughout time. The majority of efforts went into reducing the influence of noise. The minimum variance distortionless response (MVDR) [11] for example separates correlated and uncorrelated signal components to reduce the malicious effects of noise. Other beamforming techniques use filtering techniques. A drawback of these beamforming techniques, which is mentioned frequently, is their poor resolution capabilities. This becomes particularly apparent if these methods are applied in a multi-source situation or highly reverberant environment. Another category of DOA estimation methods are subspace-based techniques that make use of the subspace composition of the signals such as the multiple signal classification (MUSIC) algorithm [12]. These methods have several advantages over conventional beamforming techniques. In most cases, assuming the signal subspace is orthogonal to noise works well in most scenarios. Also, contrary to

methods that require the signals to be transformed in frequency space, MUSIC is not constraint to frequency bins resulting from block processing due to a Fourier transformation. However, subspace decomposition does not directly differentiate the signal space from the rest, hence the amount of sources is required to derive a meaningful output. While MUSIC is the most notable algorithm in this category, there are numerous adaptations. Estimation of Signal Parameters via Rational Invariance Techniques (ESPRIT) [13] for example is considerably more computationally efficient, considering the costly decomposition algorithms, but it imposes geometric constraints on the sensor array. Others claim to improve capabilities for narrow (TOPS) [14] or wideband (CSSM) [15] source signals. Another category in common DOA estimate approaches may be discovered when moving away from solely signal processing techniques. Maximum likelihood methods [16] [17] define signal models and estimator that retrieve fitting parameters. These ML approaches work well on their own, however, they are frequently reliant on certain model assumptions, to narrow the parameter search space, that often do not hold true. This limits the versatility of ML methods in comparison to their equivalents.

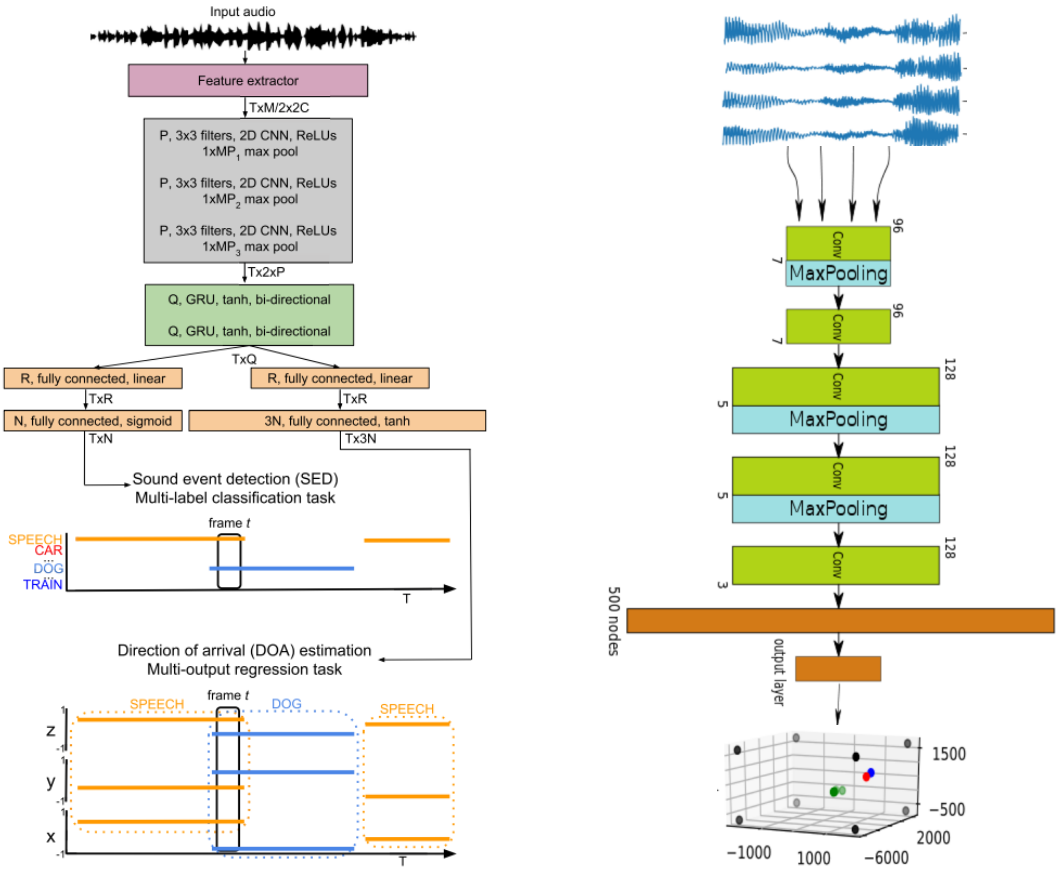
The methods stated here have shown promising results in various applications, however come with multiple flaws. A lot of critique is directed towards the general low angular resolution capabilities [18]. Furthermore, as basic signal processing techniques, the DOA output requires an additional process to distinguish an alleged source from noise or reverberation. Although some methods can evidently reduce these malicious effects, it is of the nature of those methods that the coherent signal is transformed into a vector of directions. Therefore, losing characteristics of the source, making white noise indistinguishable from speech or other audio. The signal has to undergo a bandpass filter to selectively localize different sources. This is in particular troublesome if multiple sources are to be detected and localized that are of different characteristics. Especially with this problem of signal-source association in multi-source situations, the basic algorithms reported here do not provide the flexibility for more complex tasks. In the following, those specifics are addressed and examples are shown in Figure 2.1.

2.1.2. Learning Based Techniques

Learning based techniques have recently gathered much attention in the domain of SSL. Further driven by conference workshops and challenges, numerous papers have been authored themed with acoustic classification and localization methods. The acoustic source localization and tracking (LOCATA) [19] and Detection and Classification of Acoustic Scenes and Events (DCASE) [20] are noteworthy sources of which. Not only have many innovative, application-driven approaches emerged from these workshops, but they also provided datasets that may be used as benchmarks outside of the scope of the challenge. A common concept in the domain of data-driven SSL is the use of convolutional neural networks (CNN). Many of them claim to outperform signal processing counterparts, particularly in noisy and reverberant environments. Multiple, overlapping sources are therefore also resolved much more robustly. Besides performing comparatively better, CNNs are much more versatile in the desired application. Combined sound event classification and localization have shown to be quite successful in multitask networks [21] [22] (see Figure 2.1a), which is not possible with merely signal processing techniques. With a neural network architecture, the output is also not constraint to be sampled vector of DOA angles, but can also be a continuous function [8] or extend the output to three dimensional coordinates [23] [24]. But even in the standard acoustic camera/imaging approach, a method that overlays a camera view with the heatmap of the likelihood of a sound source, much higher resolutions have been achieved with the application of deep learning [25]. Most methods rely on some form of spectral input data [23] [24], but working examples of networks feeding raw sensor data have already been reported [26]. An example of such a network is displayed in Figure 2.1b.

2.1.3. Sound Source Localization in Robotics

These and other approaches have been utilized with considerable success in robotics and vehicle technology in particular [27] [18]. A common use case is found in surveillance purposes on board of micro aerial vehicles (MAV) [28] [29]. However, due to the close proximity of the sensor to the actuating rotors, this application suffers from very low signal to noise ratios (SNR). Wheeled robots show a much better potential as a platform for acoustic techniques. Most of the research focuses still on indoor mobile platforms. This has also influenced the design of including multiple microphones in the mobile robot platform known as Pepper [30], which has enabled further research in this field [22]. Despite the fact that speaker tracking remains the most popular use case, new research has shown that active and passive acoustics may also be used for simultaneous localization and mapping (SLAM). These approaches, most notably [31] and [32], create an environment



(a) The multitask SELD-net network by Adavanne et al. [21].

(b) An end-to-end approach for SSL by Vera-Diaz et al. [26].

Figure 2.1: Examples of state of the art sound source localization methods. Figure 2.1b shows an end to end approach to the SSL task without any feature extraction. The network is trained on semi simulated data and fine tuned on real data. Figure 2.1a shows a multi task network. The method solves the task of sound event detection as well as sound source localization by diverging branches of the same network.

by emitting acoustic pulses and filter direct from echoing paths to create a representation of the robot's surroundings and localize in it. Acoustic techniques are also well represented in automotive research. While the actual use of acoustic localization is not one of them, studies addressing the decrease of road vehicle noise emissions are frequently accompanied with some technique of acoustically localizing those sources. One of the few examples, where acoustic sensing from the vehicle perspective has been the focus of the research is [33]. The methodology of this approach is based on a basic signal processing technique. However, it concentrates more on the problem's engineering and practicality, as well as mounting and integrating the sensor array flush on the chassis and ensuring that the implementation runs in real time.

These methods will not provide a useful output when dealing with obstructed sound sources. Although not stated, all require indirectly a line-of-sight towards the target. Research specifically towards acoustic non-line-of-sight detection is sparse, but some methods exist. The main difference to LOS are certain physical assumptions of the soundscape or environment. In [34] the authors use Matched Field Processing in a miniature test scenario and simulations to estimate the location of a sound source behind a rigid wedge. A physical model of the expected soundfield surrounding this wedge is derived, and the correlation between target location and array position is estimated using the array measurements. Other methods follow a similar scheme. In [5] [35] and [36] direct methods could not resolve sources that undergo diffraction or reflection effects. They use raytracing with a simultaneously mapped environment instead, successfully rendering an accurate origin of the occluded sound source. Other NLOS localization methods have benefited from similar strategies as well. To identify oncoming road users from obscured regions, [37] for example uses a reflection

aware RADAR technique. There are ways, omitting the physical model dependency from methods like these. Gannot et al. [38] propose a learned alternative of soundfield representation. The authors propose to use the relative transfer function as a characteristic of room acoustics that incorporates the physical impacts, a soundfield experiences, as it travels from source to receiver, using a manifold learning method. The learned manifold can be seen as the acoustic representation of one or more acoustic environments. With a semi-supervised approach, the authors propose to adapt the method towards new, unseen environments. While there is no evidence of using this method in NLOS situations, it could compensate the necessity of either a direct line-of-sight or further knowledge about the current environment.

2.2. Contribution

In this thesis, the problem of non-line-of-sight acoustic localization in the context of autonomous driving applications is explored. Two data-driven methods are proposed that aim to detect other road users in shadow zones of common occlusions in today's traffic landscape. The first method uses a simple classifier approach that uses direction of arrival features, that serve as a description of the surrounding soundfield. The second method elevates this concept a step further. The initial signal processing step stays the same, however, a neural network approach is injected in the feature formation process, that introduces weights for the combination of the dimensions of time, frequency and microphone signal pair. A situation in which the ego vehicle can benefit vastly from additional information in occluded areas will pose as the basis for this investigation. Therefore, a novel, real-world, audio-visual dataset featuring this particular environment is recorded. This dataset includes variance of the acoustic footprint of the surroundings and the dynamic state of the ego-vehicle. The methods are therefore not only tested on individual test samples but their generalization capacity across parameter changes are demonstrated as well. Furthermore, it is shown how much earlier the proposed methods can detect targets before they enter line-of-sight. To do that, the results over the entire time horizon of the recordings are compared to a state of the art visual detector running alongside. While it is shown that the proposed methods can detect targets significantly earlier than any conventional sensor, changes in the acoustic environment can affect the acoustic pipeline harshly. However, as of the author's knowledge, the proposed methods are the first pure acoustic localization approaches that successfully detect NLOS targets without any other cues of the environment.

To summarize, this thesis adds the following novel contributions:

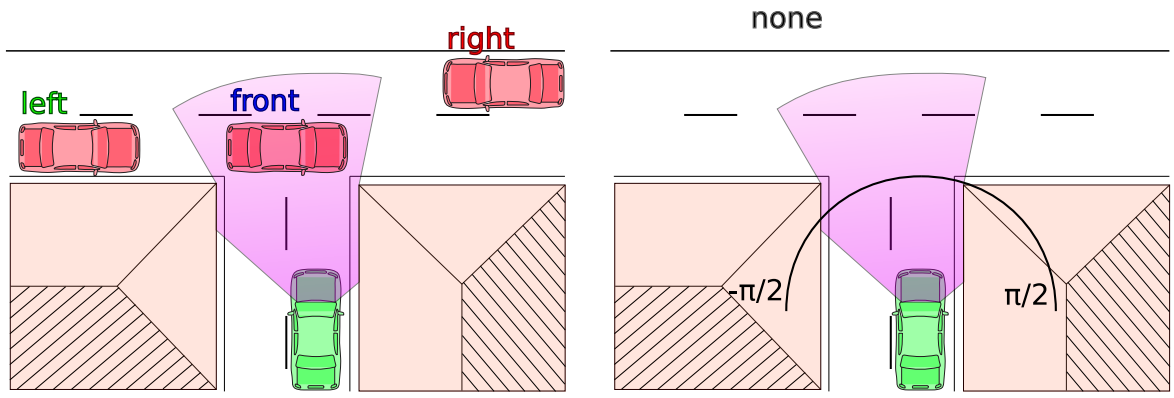
- A new, real world dataset is recorded that is comprised of different locations and variational acoustic characteristics.
- Two novel methods are proposed that can localize occluded vehicles using only acoustic cues.
- The methods are critically assessed towards the feasibility of an actual real time, real world, end to end deployment.

3

Methodology

To solve the task of detecting occluded vehicles purely by acoustic means, it is proposed to formulate the problem as a classification task. The T-crossing situation is defined as the configuration of two perpendicular road sections. As depicted in Figure 3.1, the ego vehicle is positioned on the ceasing portion while facing the continuous part. In this configuration oncoming traffic is expected to arrive from the occluded zone either to the left or right of the ego vehicle. The classifier, however, should not be limited to this binary distinction, but should also be able to tell whether there is no traffic ahead or if the target has already moved into line-of-sight. Therefore, the classification output is defined as one of four possible classes. The four classes in the set $\mathscr{C} := \{\texttt{left}, \texttt{front}, \texttt{right}, \texttt{none}\}$ are defined as follows:

- `left` denotes the event of an approaching target vehicle from behind the left corner.
- `front` denotes the event of a visible passing car in front of the ego vehicle.
- `right` denotes the event of an approaching target vehicle from behind the right corner.
- `none` denotes the event of no vehicle present in the vicinity.



(a) Displaying class `left`, `front` and `right`

(b) Displaying class `none`

Figure 3.1: Illustration of the reference situation. The purple cone shows the shadow zone of conventional sensors that require line-of-sight, missing the red target cars, coming from left and right. 3.1a illustrates the definition of classes `left`, `front`, `right` and 3.1b the class `none`.

Figure 3.1 shows an illustration of the four classes. It is worth mentioning that in the scope of this thesis the events occur mutually exclusive. There is not an incidence at which more than one of the four are true at the same time. Furthermore, the raw input shall only be the audio streams that are recorded by the vehicle-mounted microphone array. As an outdoor real world environment, the ego vehicle is subject to general

environmental noise. It is assumed that only traffic occurs in front of the ego vehicle as shown in the Figure. The ego-vehicle can also be considered mobile in this setup, where it is moving towards the trajectory of the target vehicle. The target vehicle is travelling at speeds that are appropriate to the traffic situation, i.e. around $20-40$km/h without slowing down at the intersection. At these speeds, audible signals are ensured at the position of the ego vehicle in the crossing alley [39].

3.1. Classification

There have been various attempts to tackle NLOS localization, as discussed in Section 2, but none have demonstrated substantial success in any real world scenario without assuming further knowledge about the scene or relying on a different sensor modality. While previous approaches utilizing only pure acoustic sensors succeeded in simulated settings, the fidelity of the localization goal may have contributed to the method's modest success. Recalling the example [34], Singh et al. tried to localize an occluded sound source on a grid in the shadow zone of a simple wedge, outputting a heatmap that should yield exact coordinates of the sound source. This setup is challenging for two reasons. First, the wedge only allows for diffraction effects, limiting the total number of routes at which the soundwaves can travel to reach the receiver. Second, the grid evaluation leaves the problem with a lot of degrees of freedom in the form of exact coordinates that must be estimated. The problem statement in this thesis has been greatly condensed. Only the general layout of the T-crossing, as illustrated in Figure 3.1, will be used to determine the position. In this thesis, the problem statement is significantly simplified. The only assumption about the location shall be the rough layout of the T-crossing as shown in Figure 3.1. It is further not of interested at which x or y coordinate the target vehicle is at any given point in time, but whether it is approaching and from where. To achieve this, the directivity pattern of the soundfield has to be extracted via some form that can serve as input to a classifier.

3.1.1. Directional Feature Formation

In binaural audio applications common features are the inter-aural time difference (ITD) and inter-aural level difference (ILD) [40] [18]. Both simply extract the difference of phase for the former and magnitude components for the latter from the frequency transformed signals. This method, using the ILD, draws inspiration from the human hearing mechanism, where signal strength can additionally relate to sound source location [41]. This comes from the directional characteristics of the ears that are collecting the signals from opposite directions and are further shielded by the head. The way multi channel microphone arrays are usually constructed with omnidirectional microphones on a skeletal frame, like the one used in this research, make magnitude based features not viable for this research. As a result, most acoustic localization methods with a similar setup employ a combination of phase difference and correlation approaches to infer the position of a sound source. For multiple sensor array configurations, beamforming works similar to the ITD for binaural installations. Via constructive and destructive interference of the individual receiver signals, a corresponding output magnitude relates to an angular direction of arrival (DOA) of the source signal. The phase transformed steered response power (SRP-PHAT) technique [10] is the methodology utilized in this thesis. This method also has the advantage of processing data in the frequency domain, allowing for selective frequency filtering. Furthermore, the phase transform mitigates the impacts of echoing signals, making the output more identifiable in a variety of acoustic environments. In the following an overview of the SRP-PHAT method by [10] is introduced, which is provided for completeness and to establish notation.

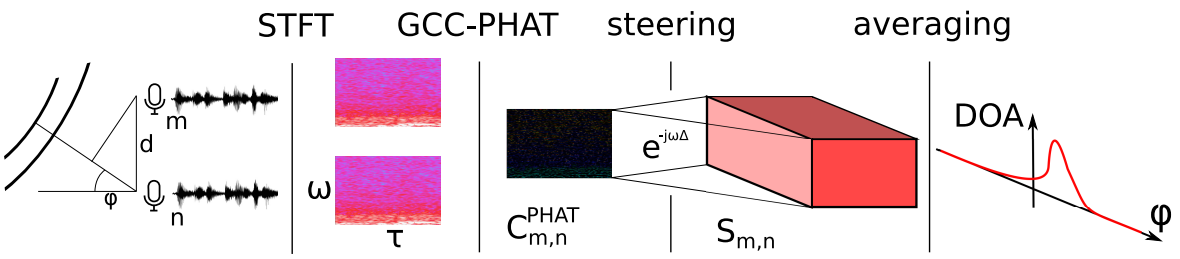


Figure 3.2: Illustration of the basic SRP-PHAT method. For simplicity only a two microphone setup is displayed. Leftmost the two microphones in the directional soundfield are displayed. The two microphones receive the phase shifted signals as a function of impingement angle ϕ. The sequential processing towards the right illustrates the formation of the DOA vector through STFT, generalized cross correlation with the phase transform (GCC-PHAT), steering and averaging.

The method requires the signal $x[t]$ to be transformed by the Short Time Fourier Transform to $X[\omega]$. The Hann window was used to calculate the transformation with an overlap of half the segment length. The STFT spectrograms of two microphones are then correlated to one another and the phase transform is applied with

$$C_{m,n}^{PHAT}[\tau,\omega] = \frac{1}{|X_m[\tau,\omega]X_n^{\star}[\tau,\omega]|} X_m[\tau,\omega]X_n^{\star}[\tau,\omega] = \Psi_{m,n}[\tau,\omega]C_{m,n}[\tau,\omega] \tag{3.1}$$

m and n denote two distinctive microphones and τ, ω the Fourier transformed sequence parameter and frequency. $\Psi_{m,n}[\tau,\omega]$ is the weighting function known as the Phase Transform, that basically whitens the spectrum of both of the signal spectrograms. This is supposed to narrow down the actual direction signal content in the cross correlation matrix $C_{m,n}[\tau,\omega]$. As part of the method, each temporal frequency bin $c_{m,n}$ at any given frequency and time (τ_x, ω_x) of the phase transformed correlation matrix $C_{m,n}^{PHAT}[\tau,\omega]$ needs to be steered with the delay of the corresponding microphones as illustrated in Figure 3.2. The delay can be expressed as a function of impingement angle which can be sampled across the region of interest as

$$\Delta[\phi] = \frac{c}{d[\phi]} \qquad d(\phi) = d\sin(\phi) \tag{3.2}$$

Δ is the steering delay that is sampled at the corresponding impingement angles ϕ and calculated via the spatial separation of the microphones d and the speed of sound $c = 343$ m/s. Each element $c_{m,n}$ forms a steered vector in the range of ϕ around the microphone pair at a particular frequency and time with

$$s_{m,n}[\phi] = c_{m,n}e^{-j\omega\Delta[\phi]} \tag{3.3}$$

Combining all steered partitions $s_{m,n}[\phi]$ results in the 3 dimensional function $S_{m,n}[\phi,\tau,\omega]$. The DOA vector of the SRP-PHAT method can now be achieved by simply averaging across the frequency dimension of $S_{m,n}[\phi,\tau,\omega]$. At a particular sequence step τ_x and only considering the microphone pair (m,n), the DOA vector reads for all F frequency bins

$$DOA[\phi] = \frac{1}{F}\sum_{\omega} S_{m,n,\tau_x}[\phi,\omega] \tag{3.4}$$

This process is generally extended across the time horizon of interest (sample length) and spatially according to physical capabilities of the sensor array. For more than two microphones, the same process is applied for each available pair out of the number of individual microphones M. The time horizon or sequence length T is determined by the sample length δt and the STFT parameters. The number of available microphone pairs in relation to the number of microphones is $\hat{M} = \binom{M}{2} = M(M-1)/2$. Both are averaged over the respective dimension. So for all microphones, snapshots and frequencies the total DOA vector can be expressed as

$$DOA(\phi) = \frac{1}{\hat{M}FT}\sum_{m}^{M}\sum_{n>m}^{M}\sum_{\tau}^{T}\sum_{\omega} S_{m,n}(\phi,\omega,\tau) \tag{3.5}$$

The DOA output is affected by the grid on which it is sampled on. The grid used here is the azimuth direction around the ego vehicle as displayed in Figure 3.1b. The range is limited to $-\pi/2$ to $\pi/2$, centered around the driving direction. This captures soundwaves impinging on the vehicle in the horizontal plane, resulting in a vector corresponding to the possible angles in the provided range. In most LOS applications this vector is enough to detect and successfully localize sound sources by determining the peak power and hence the DOA of a particular sound source. Because a specific direction is not available in NLOS scenarios, the DOA vector is inconclusive. However, still providing a signature corresponding to the current soundfield, will provide the cornerstone of the proposed classification methods.

3.1.2. Uniformly Weighted Support Vector Machine

The first proposed method, further referred to as Uniformly Weighted Support Vector Machine (UW-SVM), is based on the conventional SRP-PHAT method, in which the resulting tensor is averaged across all dimensions (F, T, $\hat{M}$). Before doing so, some modifications are proposed to tailor the standard DOA vector to the situation at hand. Recalling the setup in the beginning of this chapter, the situation is dynamic by nature, where the target vehicle is in motion. The standard SRP-PHAT method extracts a snapshot of the data it had been fed by averaging over all time sequences of the matrix S. Feeding a very long time period can squash the data and feeding a very short one can amplify noise. In order to capture dynamic changes of the soundfield, as

well as maintain a meaningful snapshot, it is proposed to split the data into multiple segments L, that capture snapshots with a temporal offset to one another. By including a rate of change in the feature vector like this, it is expected that the suggested method's accuracy would improve. The DOA vectors of the snapshots are concatenated to form a single feature vector.

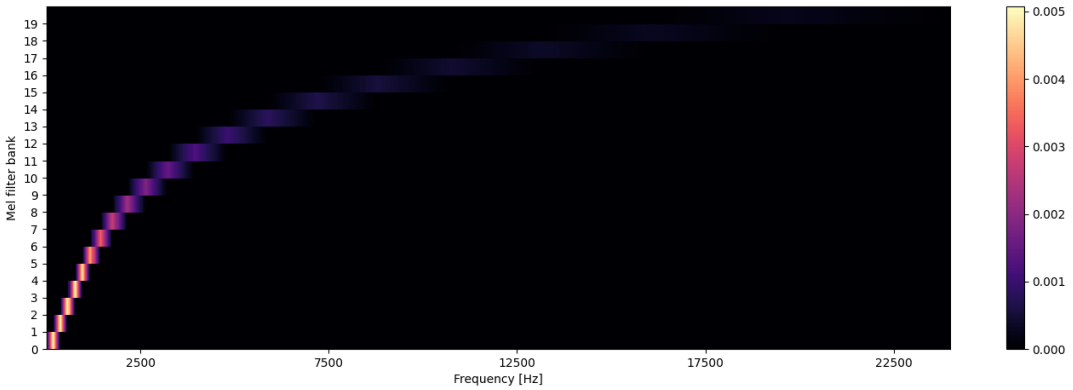


Figure 3.3: Illustration of the transformation matrix mapping onto the Mel filter banks for 20 filters. The mapping shows linear characteristics in the frequency range of $0-2000$Hz. For frequencies >2000Hz the plot shows the non linear relationship as described here.

Frequency manipulations were also explored as a way to enhance the feature vector. Many techniques showed better results utilizing the signal on the log Mel filter banks rather than using simply the raw STFT spectrograms [23], [24], [42], [43], therefore it can be considered as a common strategy in comparable research. The Mel scale is the result of a non-linear transformation of the frequency space. It is based on the fact that humans perceive the tonal difference of a frequency band of lower frequencies much better than an equidistant band in higher frequencies. For example, the difference between 50Hz and 100Hz is perceived much more distinct than the difference between 10050Hz and 10100Hz. The assignment of frequency regions to filter banks for the Mel transformation is plotted for 20 filters in Figure 3.3. Despite the fact that this mapping is based on differences in human perception, numerous research has found that utilizing this nonlinear mapping as a building block in their approaches for both SED and SEL tasks, with no constraints on human voice sources, has a significant effect. However, the methods presented here only localize targets from the universal tyre noise of motorized road vehicles. As multiple studies have shown, tyre noise has a broad, symmetrical frequency spectrum, which is centered around 1000Hz [44] (also see beginning of Chapter 2). Considering the region of interest for tyre noise, it can be clearly seen that the Mel filter banks in the region of $[0,2000]$Hz are linearly distributed (see Figure 3.3) and hence would yield the same result with or without the Mel scale mapping. Hence, for this approach, a simple band-pass filter is used instead, which cuts off unwanted frequency ranges. Doing so, the matrix $S_{m,n}$ is simply divided, and only frequency bins of interest are further processed as shown in Figure 3.4. It is worth mentioning that the band-pass may be applied directly after extracting the STFT spectrograms, saving computation of the non relevant frequency bins. For illustration purposes, the split is displayed for the $S_{m,n}$ matrix.

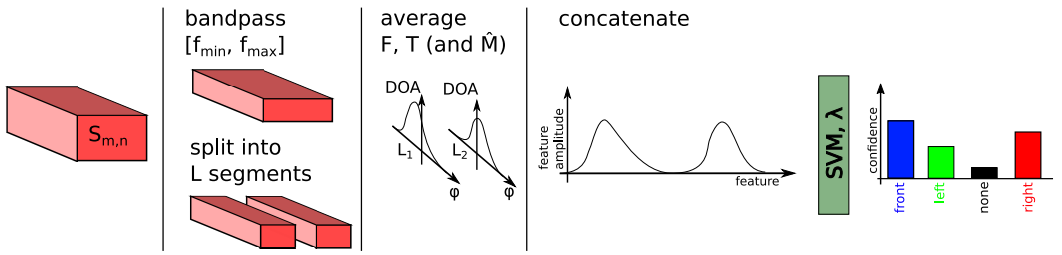


Figure 3.4: Illustrating the UW-SVM method. For simplicity, only a two microphone setup is displayed. First, the matrix $S_{m,n}$ is split into two separate snapshots and the bandpass filter is applied. The averaging follows the standard SRP-PHAT technique shown in Figure 3.2. The DOA of both snapshots are then concatenated and fed into the SVM classifier that produces the infered labels.

With the feature formation rooted in a generic DOA extraction, some augmentations are possible. Assuming a sound source that is situated at $\phi = 45°$ in the azimuth direction in front of the sensor array, the SRP-PHAT DOA vector would take a specific shape. It can be assumed that an equivalent sound source situated at $\phi = -45°$ in front of the sensor array would produce the same shape, only mirrored at the ordinate of the coordinate system. With the configuration shown in Figure 3.1, it can be assumed that the problem description is symmetric enough that mirroring the DOA vector derived from the audio samples in this context would provide equivalent results as if the samples from the classes `left` and `right` were reversed. By selectively flipping the DOA vectors and assigning the opposite ground truth labels, it is feasible to supplement data including `left` and `right` samples to obtain double the quantity. This must be done before concatenating the DOA vectors as illustrated in Figure 3.4.

Following the feature formation described here, the parameters that will be used are $L = 2$ for the number of segments of a $\delta t = 1$s audio sample and $f_{min} = 50$Hz and $f_{max} = 1500$Hz for the band-pass cut-off frequencies. Furthermore, the grid resolution of candidate angles has to be defined before sampling. For the defined range of $[-\pi/2, \pi/2]$ a grid size of $B = 30$ is chosen for the DOA vector, rendering the size of the feature vector to $2B = 60$ bins. This resolves to the feature vector of the UW-SVM method to

$$\boldsymbol{u} = \left(DOA_{L1}(\phi)\big|_{-\pi/2}^{\pi/2}, \quad DOA_{L2}(\phi)\big|_{-\pi/2}^{\pi/2} \right)^\top \tag{3.6}$$

The feature vector is then fed into a basic support vector machine (SVM). The SVM is handled in a one versus one fashion for the multi class problem that outputs affiliation to one of the class lables. The ℓ_2 regularization parameter λ is set to 1 and the squared hinge loss is used to optimize the parameter set $\boldsymbol{w}$ with the loss function

$$\min_{\boldsymbol{w} \in \mathbb{R}^{L \cdot B}} \sum_{i}^{N} \left(\max\left(0, 1 - y_i \boldsymbol{w}^\top \mathbf{u}_i\right) \right)^2 + \lambda \| \boldsymbol{w} \|_2^2 \tag{3.7}$$

Here, y_i is the label of sample i from the set $\mathscr{C}$ and N the number of all training samples. Also, to enable confidence values for the SVM classifier, Platt scaling is employed [45]. This method fits a scaling on the numeric estimates of the SVM output by a cross correlation on the training data.

3.1.3. Selectively Weighted Convolutional Neural Network

Ideally, the signal processing step can be avoided by directly feeding the raw audio wave samples to a neural network, that is itself capable of learning how to extract the phase offset between the audio lines and infer to one of the classes defined in Section 3. However, the relationship between raw audio samples of multiple microphones to the desired output can be considered highly complex and non-linear. A more complex relationship warrants more complex models that require more data. Because the entire dataset is recorded supplementary to this thesis, the data size is limited. Therefore, the second method is only extended to include the averaging step that forms the DOA vector in the standard SRP-PHAT method. The goal is to train a network that learns a selective weighting to include different components in the final feature vector, rather than squash all components in the S tensor that may or may not be relevant by averaging. To do this, the feature formation step is included in the classification method itself. A convolutional neural network is employed to replace this process and optimize the weighted parameters on a training set. Therefore, this method is referred to as Selectively Weighted Convolutional Neural Network (SW-CNN).

The input tensor S to the network is defined as the 4 dimensional function $S[\omega, \tau, \phi, \hat{m}]$ that is combined from Equation 3.3 with dimensions (F, T, B, $\hat{M}$). Real and imaginary parts are separated and concatenated in the dimension of the steering angle to accommodate the complex data. Each dimension is successively filtered by a one dimensional kernel of the same size as the corresponding dimension and convoluted across the other. This imitates the averaging process from Equation 3.5 in a learned fashion. For example, the first convolution step operates a kernel of size F. The convolution runs zero steps across the frequency dimension, but $\hat{M}$ steps across the dimension of the microphone pairs and all other orthogonal dimensions like

$$S[\omega, \tau, \phi_{Im,Re}, \hat{m}] * K[\omega] = \sum_{\bar{\omega}} S[\omega - \bar{\omega}, \tau, \phi_{Im,Re}, \hat{m}] K[\bar{\omega}] \qquad \forall \tau, \phi_{Im,Re}, \hat{m} \tag{3.8}$$

Here, $\phi_{Im,Re}$ denotes the variable of the steering angle that is doubled with the imaginary and real part of the input tensor. $\hat{m}$ is the variable of the microphone pairs and $\bar{\omega}$ is the convolution shift variable. This

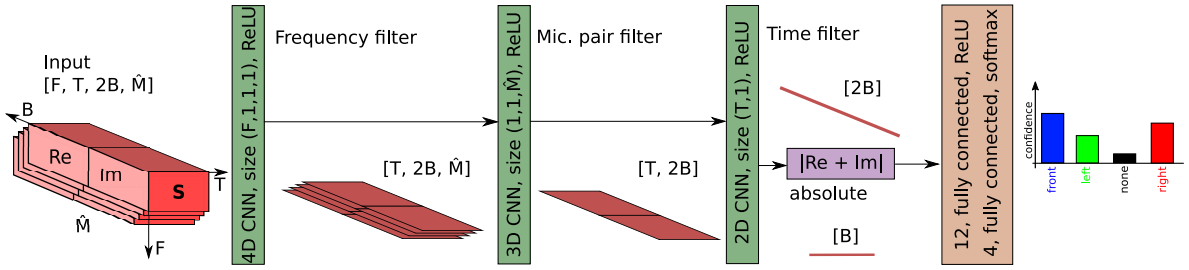


Figure 3.5: Illustration of the SW-CNN method. In the graphical representation the 4th dimension ($\hat{M}$) is depicted by stacking the tensor above one another. The green layers show the convolutional layers that incorporate the weights of the selective averaging procedure. The purple layer denotes the parameterless arithmetic operation of merging the complex number. The orange layer displays the two fully connected layers.

convolution with the kernel K is shifted over the other the number of dimensions determined by the input. Each layer uses a ReLU activation function as it is displayed in Figure 3.5 and the CNN and tensor dimension reduces each layer. After the weighted averaging the reduced tensor of size $2B$ is merged with itself to combine the real and imaginary parts, that were split for the input. The merging is simply defined by taking the absolute value, reducing the layer further to the size B. Therefore, this intermediate feature can be seen as a learned representation of the first method. Lastly, two fully connected layers produce the output with a softmax activation function for the four classes.

Despite using a learned representation of the feature formation, the 4 dimensional input tensor increases in size so much, that handling that much data becomes unpractical. As a result, the input tensor is trimmed similar to the frequency filter of the UW-SVM method. By means of a temporary ℓ_1 regularization, all convolutional layers are separately trained and analyzed. If a feature bin in any dimension expresses significantly less contribution than others it will be cut before the input formation for the final approach (see Section 4.2.2 for details). Based on these findings all non-zero frequency bins of the kernel have shown to be present in the range of f_{min} = 0Hz and f_{max} = 1900Hz. All non-zero microphone contributions resolve to 11 distinct microphones. The time sequence kernel showed several zero instances, however, spread across the entire sample length. To preserve the susceptive field, every third time instance was used.

3.2. Data Acquisition

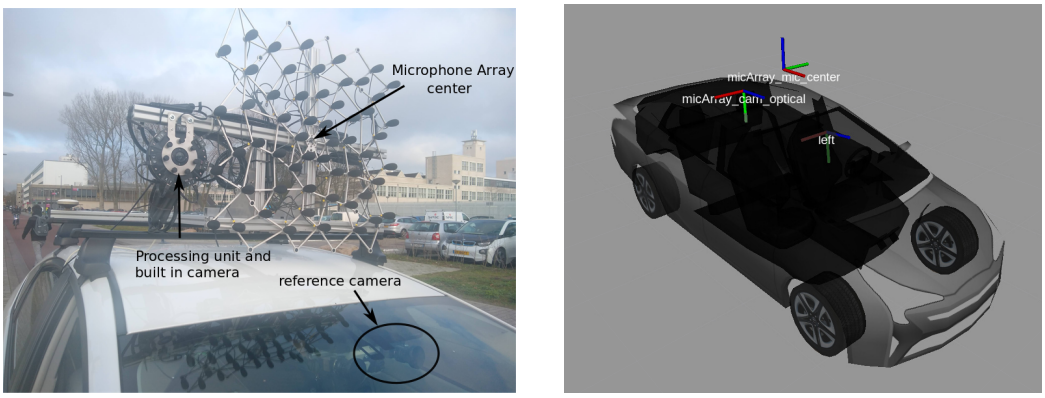
Sound source localization tasks have a wide range of applications and this variance has led to a lack of generalizing datasets. Many approaches solve niche issues, similar to the topic of this thesis. As a result, new concepts are either bound to be conducted in simulated environments or require extensive work to generate individual datasets. Consequently, there is no dataset to fall back to for the investigation of non-line-of-sight localization in traffic scenarios. A simulated environment was ruled out for a variety of reasons. First, the fidelity of a real world environment cannot be reproduced in full capacity, which can capture all reflection and diffraction phenomena that this method relies on. Second, capturing vehicle noise is difficult since the target is in motion, which is intrinsically responsible for sound creation as tyre-road interaction, making its usage questionable at best and recording impractical. Therefore, a new dataset is recorded, tailored to facilitate the research as described in Section 1.2.

3.2.1. Vehicle Platform

With the intention of performing real environment recordings, the research vehicle of the Intelligent Vehicles Group of the TU Delft was used as the operating platform and interface for the novel sensor array [46]. The converted Prius (see Figure 1.2 for a reference picture of the Prius) is equipped with a steady DC supply from the internal battery, powering a compute unit that processes the sensor input of the local network. The platform is equipped with an array of different sensor technologies aiding other research fields within the group, however, for the scope of this thesis, only two are of importance. A front facing stereo camera and the microphone sensor array.

Stereo Camera:
The front facing camera rig consists of two Ueye cameras (UI-3060CP-C-HQ R2) with a baseline of 22cm. The software side trigger is limited to a constant 10Hz for the entirety of the data acquisition, if not stated otherwise. The purpose of the rig will be constraint as an extrinsic anchor to the vehicle chassis in the virtual environment, labeling and validation of results and baseline approach (see Chapter 4). Therefore, only the left camera and perspective is used as indicated in Figure 3.6a.

Microphone Array:
The research microphone array by cae-systems is based on an integrated acoustic camera. It is made up of 56 microphones, that are randomly distributed on a 0.4m × 0.5m aluminium web structure. Each 12mm ADMP441 MEMS microphone connects to the processing unit in bundles of 8 via a network cable. The processing unit also houses an additional low resolution camera, originally designed to support visualization of the built-in processing methods. Both are rigidly mounted on an aluminium beam structure on the same vertical plane. To further reduce external effects, vibrations in particular, rubber spacers are put in between the web and beam structure.



(a) Sensor placement and attachment on the actual vehicle.

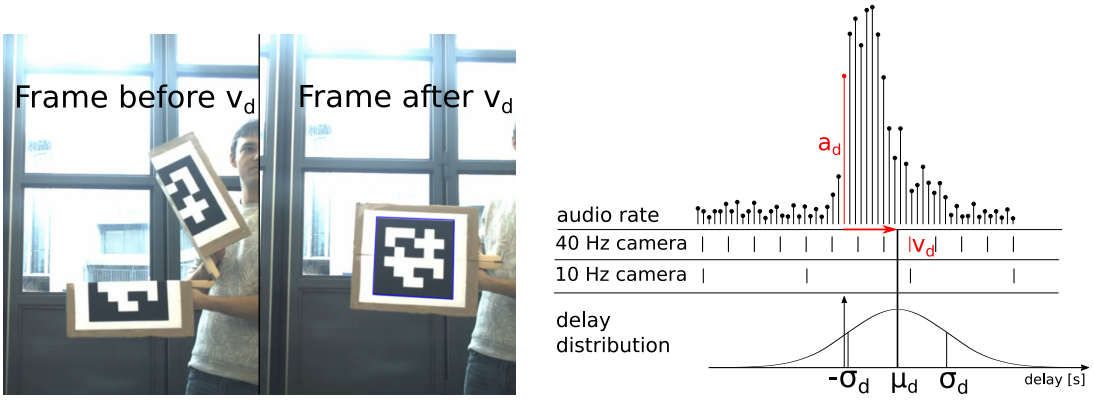
(b) The virtual environment and reference frames of the sensors after external calibration.

Figure 3.6: Displays of the sensors that are used on the vehicle. 3.6a shows the sensors on the real car, while 3.6b displays the reference frames in the virtual environment.

3.2.2. Data Processing

Running on the vehicle side compute unit is an Ubuntu 18.04 system with the Robot Operating System (ROS), allowing communication and processing of the sensory inputs. Because the microphone array is running on a separate integrated circuit it takes over clocking for the data acquisition. Other sensors run directly on the ROS-PC, therefore, it is expected that sensor messages between the camera and audio signals are not synchronized. While the processing unit ran proprietary software from cae-systems, it was not possible to employ an out of the box time synchronization protocol directly on the hardware. Instead, the delays are analyzed for a number of samples to extract any systematic delay that can be applied while extracting the data to standalone sound and video files.

For the analysis of time delays, an audio-visual event was created by attaching a single April tag [47] on a starting clapper. A printout of the tag was cut in half and fixed on the clapper as shown in Figure 3.7a. The camera will be able to detect the frame v_d, in which the full, combined tag, is visible. The sound pressure peak from a forceful clap can then be used to identify the moment of the sound event a_d. Of course, this technique restricts the resolution and hence the precision of the synchronization. However, as the camera view is used for labeling LOS and NLOS situations, it is sufficient to use the temporal resolution of the camera and synchronize the microphone array with it. Even though, for this analysis, the camera frame rate is increased to 40Hz. The most essential thing is to avoid a positive shift, which occurs when the clap triggers the microphone before the tag is detected in camera view. A sample set of 56 was recorded, audio and video line, and v_d, a_d were extracted. Since the actual clap is expected to happen between the frame, the tag was detected in the video, an offset of half the frame rate is considered in the calculations. That means, the extracted delay per sample is $\delta t_d = v_d - 0.0125 - a_d$. Taking the average over all samples, it is estimated that



(a) Two frames from the an arbitrary sample showing the clapper before and after v_d

(b) Visualization of the delay estimation. The coordinate frame is symbolically aligned with a_d. The red arrow shows the estimation of the delay in the samples. Note that the arrow reaches the 0.0125s before v_d.

Figure 3.7: A visualization of the delay estimation process. 3.7a shows the audio-visual event produced by a starting clapper. 3.7b shows

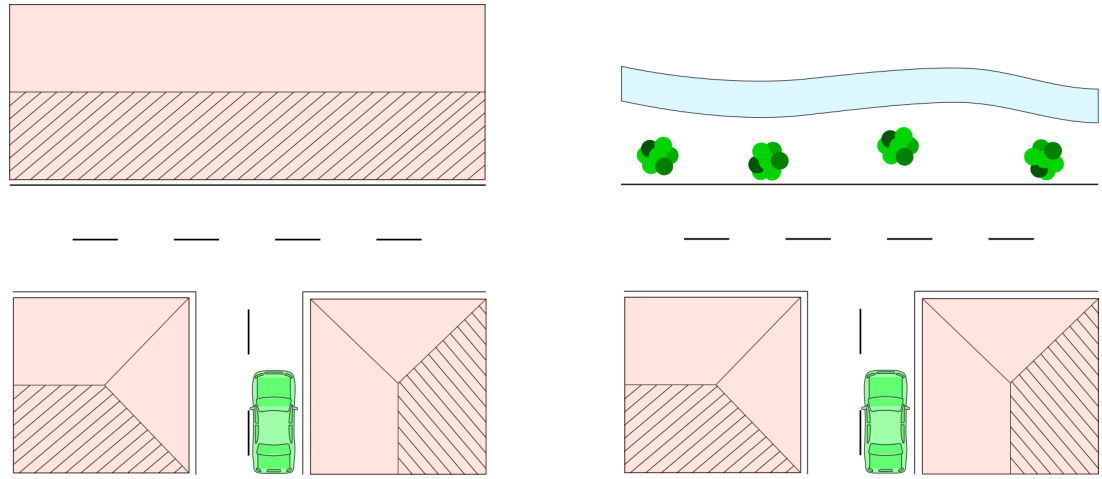
the microphone array processing unit is running on average $\mu_d = 0.046$s behind the camera, with a standard deviation of $\sigma_d = 0.039$s, which results to about half a frame. The effect of the delay between the used 40Hz camera framerate in relation to the 10Hz that is used for the experiments is illustrated in Figure 3.7b. This offset is included in the extraction of the data.

Following the temporal alignment of the sensors, the extrinsic geometric alignment of the two sensors must be determined. As the microphone array comes with a built-in camera that is fixed to a metal frame, the offset between the array center and optical axis of the processing unit can be measured directly on the rig. The stereo camera, on the other hand, is positioned inside the car and is therefore less accessible. In order to get the transformation between the vehicle mounted stereo rig and microphone sensor, an extrinsic calibration procedure was used to identify the transformation between the vehicle and microphone camera. The calibration scheme used is part of the Kalibr toolbox [48], which performs a full intrinsic and extrinsic camera calibration of any number of cameras, applying a Gauss-Newton optimization on the reprojection error over multiple cameras. It has to be noted, that due to the temporal misalignment discussed previously, the extrinsic calibration cannot recover exact results due to the different clocking. However, a rough estimation of alignment is sufficient because it is only used for visualization purposes later in Section 4.1. A steady approach with the calibration target produced visually valid results and the intrinsic camera parameters of the stereo unit matched the previously calibrated set. The sensor frames in the virtual environment are displayed in Figure 3.6b.

3.2.3. Dataset Recording

The T-crossing environment is now further divided into two categories, type A and type B. A type A location is characterized by a wall across the ego vehicle as shown in Figure 3.8a. A type B location features no acoustically effective surface, but mostly empty space (see Figure 3.8b). Acoustically effective surfaces are referred to as large and rigid walls that tend to reflect soundwaves well. Whereas trees, bushes and organic structures, in general, tend to absorb soundwaves via scattering and are not considered as such. The data was recorded in 5 different, selected locations in Delft, the Netherlands. According to the definition of environments A and B, two of which can be classified as A and three of which as B. For further identification, the environment class is followed by an enumeration indicating the location alongside (see Table 3.1).

The recording session of each location was divided into two parts: static and dynamic. In the first part, the ego vehicle was positioned inside the leading road towards the T-junction, roughly 5m from the edge of the occluding obstacle, as shown in the illustrations of Figure 3.8. The target vehicles in the recordings are from a wide range of motorized vehicles passing by, including scooters and occasional vans. To keep a steady flow, another vehicle was used, driving back and forth as long as no other road user was in sight. Therefore, a majority of the recordings include a 2010 Skoda Fabia 1.2 TSI. The samples were recorded, as far as it was



(a) Environment A; Defined by the wall across the green ego vehicle.

(b) Environment B; Defined by the lack of a reflective surface across the ego vehicle

Figure 3.8: Illustrating the two environments. 3.8a shows the completely walled off environment, while 3.8b shows the environment with an open area across the ego vehicle.

possible, alternating, i.e. one sample includes a vehicle passing from the left, then a sample where none was in proximity, the next where a vehicle was passing from the right and none again. This pattern streamlined the process of sample gathering and removed possible correlations of surrounding environment noise. If `none` recordings were conducted in a consecutive fashion instead, uncorrelated noise could not be insured. Each recording was started a few seconds before the car entered line-of-sight and stopped after it vanished behind the opposite occlusion. In the dynamic part target and ego vehicle were moving at the same time. For safety reasons, no passing vehicles were included in these recordings. Ego vehicle and target vehicle both were starting around 30 to 50 meters before the T-crossing and carefully coordinated so that they would meet closely at the intersection. The ego vehicle would stop right before the path of the target vehicle, while the latter would continue to drive past. These two cases are further referred to as static and dynamic scenarios and labeled with the letter S or D before the location identifier in Table 3.1.

Location	Recording Date static / dynamic	Abbreviation static / dynamic	Number of recordings `left`	`none`	`right`
Anna Boogerd	12.12.2019 / 11.08.2020	SA1 / DA1	33	67	35
Kwekerijstraat	16.01.2020 / 16.01.2020	SA2 / DA2	29	62	27
Willem Dreeslaan	12.12.2019 / 11.08.2020	SB1 / DB1	35	67	42
Vermeerstraat	16.01.2020 / 16.01.2020	SB2 / DB2	38	65	39
Geerboogerd	12.12.2019 / 11.08.2020	SB3 / DB3	41	81	42

Table 3.1: Overview of the recorded dataset.

Postprocessing of the data, including the labeling process, was done in a semi-automatic fashion. In a first sweep, an object detector extracts the first frame of each recording in which a vehicle was detected confidently and saves the past 15 frames to storage. Depending on the bounding box position in the frame, a class is assigned to either `left`, `right` or `none`. By visually inspecting the past 15 frames, the frame in which the target vehicle is still completely occluded is taken as the reference time t_0, splitting the recordings into distinctive periods of complete occlusion and partly and not occluded target vehicle. More specifically, the time period $t \leq t_0$ is considered as non-line-of-sight where no frame shows any portion of the target vehicle, while the rest is considered line-of-sight. For `none` cases, t_0 was arbitrarily chosen within the recording duration, a few frames before the end of the file. The definition of t_0 however cannot be applied for the dynamic data recordings. The problem emerges because the timing of the dynamic event could not be aligned perfectly. The target vehicle arrived slightly earlier or later depending on the sample. Furthermore, as the ego vehicle is moving towards the junction, the camera FOV moves along with it, sometimes cropping the edges of the wall when the target vehicle enters line-of-sight. Therefore, another reference time $\widetilde{t}_0$ is introduced specifically for the dynamic recordings. $\widetilde{t}_0$ is defined as the time of image frame that resembles the FOV of the corresponding

static event. Despite the possibility that the target vehicle is already in line-of-sight at $\widetilde{t}_0$, the static FOV, and hence the vehicle position, is far enough along the road that this is rarely the case. After building a database of recording ID, environment type, class and t_0 for all recordings, the data was cut around t_0, leaving seven seconds prior and three seconds after t_0 to limit the data size. This procedure was applied to the audio and video line, as well as the visual object detector for future reference (see Section 4.6. The latter is only true for the static recordings because the object detector did not yield reliable results as soon as the target vehicle appeared too close to the ego vehicle.

The audio sample set for the classification task can be extracted from this audio-visual database. Considering a sample length of δt, the class set $\mathscr{C}$ is extracted relative to the labeled reference times t_0 and $\widetilde{t}_0$. In the case of the classes `left`, `right` and `none` in the static recordings, the extracted time period is $[t_0 - \delta t,\ t_0]$. For dynamic recordings it needs to be ensured that the time window averages around $\widetilde{t}_0$. Therefore, the beginning and end of the dynamic samples are shifted by half the sample duration, i.e. the time period $\left[\widetilde{t}_0 - \delta t/2,\ \widetilde{t}_0 + \delta t/2\right]$. This is again true for `left`, `right` and `none` classes. For the `front` class a static offset is introduced for both occasions. To insure that the target vehicle is in direct line-of-sight, the offset was defined as 1.5s. Therefore, the time periods for the `front` classes are $[t_0 + 0.5s,\ t_0 + 1.5s]$ in the static case and $\left[\widetilde{t}_0 + 0.5s,\ \widetilde{t}_0 + 1.5s\right]$ in the dynamic case. The `front` classes are only extracted from `left` and `right` recordings for obvious reasons. That results in the amount of `front` samples shown in Table 3.2, as the sum of the equivalent `left` and `right` amounts. It is also worth to mention, that the position of the target vehicle in the `front` samples can be arbitrary. As the traffic situation varies by location, different vehicle speeds are expected and hence different positions of the target vehicle in the camera frame.

ID	`left`	`front`	`right`	`none`	Sum
SA1 / DA1	14 / 19	30 / 38	16 / 19	30 / 37	90/113
SA2 / DA2	22 / 7	41 / 15	19 / 8	49 / 13	131/ 43
SB1 / DB1	17 / 18	41 / 36	24 / 18	32 / 35	114/107
SB2 / DB2	28 / 10	55 / 22	27 / 12	43 / 22	153/ 66
SB3 / DB3	22 / 19	45 / 38	23 / 19	45 / 36	135/112
SAB / DAB	103/ 73	212/149	109/ 76	199/143	623/441

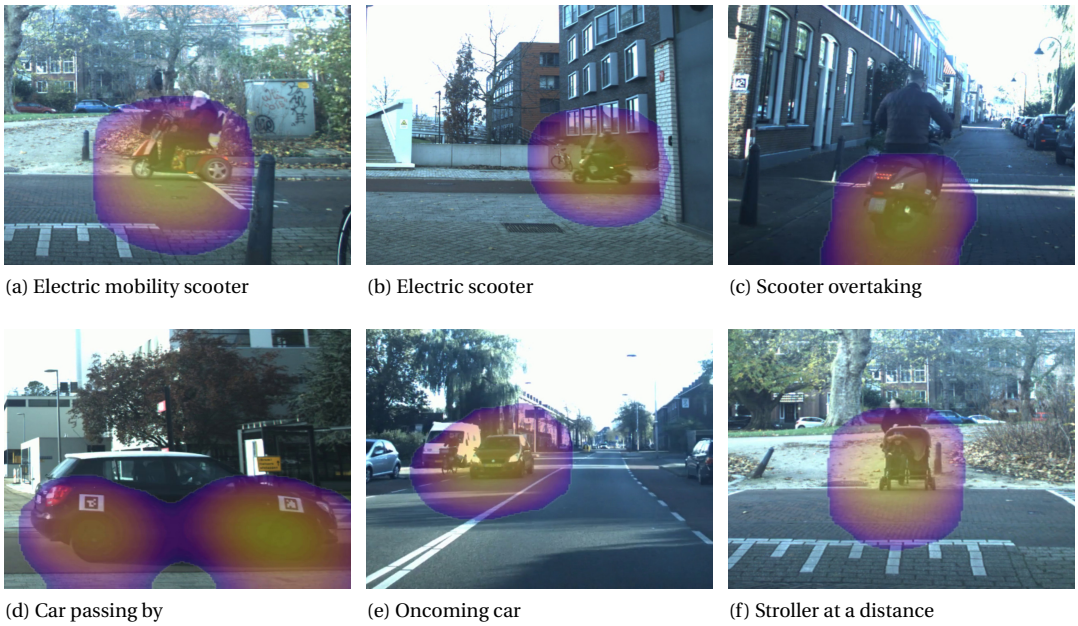
Table 3.2: Samples per subset. In the ID, S/D indicates Static/Dynamic ego-vehicle, A/B the environment type.

4

Experiments

In the following section, the proposed methods are evaluated under different aspects. Before reporting the performance on the overall dataset, the choice of hyperparameters in each method are validated. The features, learned and averaged, are compared and plotted against each other. The generalization capabilities of the methods are shown twofold. First, the generalization across the environments is analyzed. Training and testing splits are chosen exclusively based on their category defined in Section 3.2. Second, the generalization across a time horizon is evaluated. Herein, the classifier will be trained on a fixed set and the inference over multiple timesteps of the initial recordings are evaluated.

4.1. Qualitative Results and Metrics



(a) Electric mobility scooter (b) Electric scooter (c) Scooter overtaking

(d) Car passing by (e) Oncoming car (f) Stroller at a distance

Figure 4.1: Qualitative examples of 2D Direction-of-Arrival estimation overlaid on the camera image (zoomed). (a), (b), (c): Both conventional and more quiet electric scooters are detected. (d): The loudest sound of a passing vehicle is typically the road contact of the individual tires. (e): Even when the ego-vehicle drives at ~ 30 km/h, oncoming moving vehicles are still registered as salient sound sources. (f): Stroller wheels are picked up even at a distance.

The potential and relevance of acoustic localization are discussed here before moving on to the evaluation of the offered approaches. By applying a basic 2D DAS beamforming method on various video feeds that have been recorded separately to the dataset, a multitude of acoustic events could be processed. The beamforming methods evaluates the DOA angles on a meshgrid of azimuth and colatitude angles. With the transformation

matrix from Section 3.2.2 between the microphone array and the vehicle camera system and a projection built from the calibration parameters of the camera, a heatmap of the DOA grid can be overlaid with the camera frame. Figure 4.1 shows a subset of these acoustic event that were selected. Most notably, Figure 4.1b shows the results of a passing electric scooter, therefore emitting minimal engine noise, while still producing a decisive peak in its direction. This supports the assertion that, under city speed limits, vehicle noise is invariant of the internal engine, as stated previously. The same significance is held by Figure 4.1e, where a video feed was selected in which the ego vehicle drove through regular traffic at appropriate limits. While heavily subject to noise through the relative wind speed, a comparable sharp peak can be read from the heatmap, showing excellent robustness in the presence of noise.

Throughout this section, performance is evaluated as the accuracy of a classification task. The accuracy is split between an overall accuracy metric and the per class intersection over union (IoU), also known as the Jaccard index. First, for each class label y the True Positives/Negatives (TP_y/TN_y), and False Positives/Negatives (FP_y/FN_y) are computed, treating target class c as positive and the other three classes jointly as negative. Given the total number of test samples N, the overall accuracy is then

$$Accuracy = \frac{\left(\sum_{y \in \mathscr{C}} TP_y\right)}{N} \tag{4.1}$$

and the per-class IoU is

$$IoU_y = \frac{TP_y}{\left(TP_y + FP_y + FN_y\right)} \tag{4.2}$$

4.2. Concept Validation

The choices that were made in the design of the two methods in Section 3.1 have to be reasoned for. For each different techniques are applied. For the UW-SVM method a variational study is shown at which the defined hyperparameters are changed. For the SW-CNN method, a reduced network is trained in order to uncover irrelevant features to reduce the size of the input tensor.

4.2.1. UW-SVM Hyperparameter Validation

The technique used for the UW-SVM method is an ablation study. Each hyperparameter is therefore modified one by one and the accuracy of each change is reported. The parameters of interest are the inclusion of the data augmentation, sample length, number of segments and the regularization parameter of the classifier λ. The results are shown in Table 4.1 with the proposed reference method in the top row. All evaluations are cross-validated on the same folds. This includes the experiment without data augmentation. As the data augmentation samples could be correlated to already seen data, this step was performed after defining and dividing the folds and only included in the training data. Therefore, despite elevated sample size in the training including data augmentation, the folds stay the same.

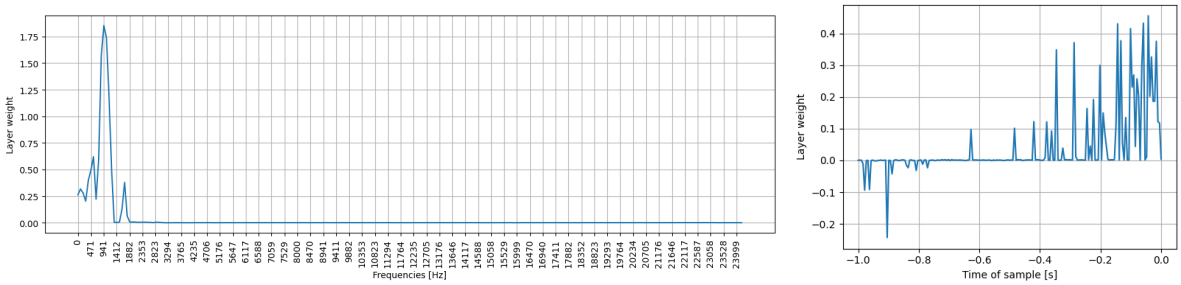
Run	Accuracy	$IoU_{\texttt{left}}$	$IoU_{\texttt{front}}$	$IoU_{\texttt{right}}$	$IoU_{\texttt{none}}$
(reference) UW-SVM	**0.92**	0.79	0.89	**0.87**	0.83
without data augmentation	**0.92**	0.75	0.91	0.78	0.83
with $\delta t = 0.5$s	0.91	0.75	0.89	**0.87**	0.82
with $L = 1$	0.86	0.64	0.87	0.73	0.79
with $L = 3$	**0.92**	0.74	0.92	0.82	0.81
with $L = 4$	0.90	0.72	0.90	0.77	0.83
with SVM $\lambda = 0.1$	0.91	0.78	0.89	0.81	0.82
with SVM $\lambda = 10$	0.91	**0.81**	0.86	0.84	0.83

Table 4.1: Baseline comparison and hyperparameter study w.r.t. our reference configuration: SVM $\lambda = 1$, $\delta t = 1$, $L = 2$, data augmentation. Results on Static data.

4.2.2. SW-CNN Feature Selection

As described in Section 3.1.3 the application of the network architecture will work on the steered and correlated tensor S of the microphone array signal. That means the input tensor is of shape $\hat{M} \times T \times F \times B$; with $\hat{M} = 1540$, $T = 189$, $F = 257$ and $B = 30$. The contribution of each element per dimension is evaluated to

filter out clutter that is not useful for the inference in order to reduce this huge input parameter space. This is accomplished by selectively applying the weighted kernel on the respective dimension and applying the ℓ_1 regularization to reduce unwanted elements to zero. The weights of interest are initialized with ones and the other dimensions are kept unweighted and are averaged out as in the original SRP-PHAT fashion. This reduced shape is consecutively fed into the same tail of the network. Each run is terminated after 600 epochs and the resulting kernel weights are plotted.



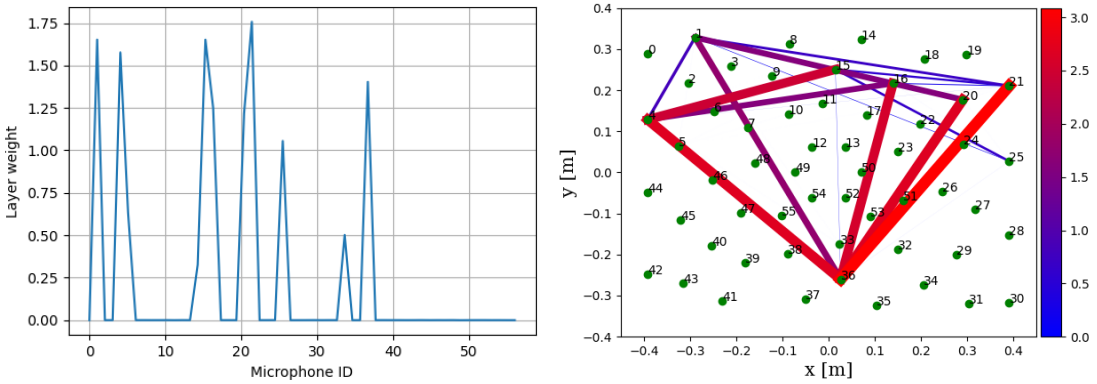
(a) Weight contribution of frequency bins

(b) Weight contribution of time bins

Figure 4.2: Plotting of time and frequency bin kernels. Figure 4.2a shows the kernel referenced in Equation 3.8, Figure 4.2b an equivalent Kernel applied on the time dimension.

Beginning with the frequency dimension, Figure 4.2a shows the clear tendency for lower frequencies, while the majority of frequency bins higher than 2000Hz are negligible.

Looking at the effect of the time bin reduction, no clear statement can be made. Figure 4.2b shows a distinct profile, where earlier frames contribute negatively and later frames contribute positively. However, there is no bandwidth of continuous zero components that warrant an exclusion from the system as it does for the frequency range with a distinct cut-off point. To preserve the perceptive range from earlier time steps to later ones, only every third bin is considered for the actual method. Also, while the training with the frequency filter active, converged fast and the accuracy reached high values as expected, the results showed a high tendency toward overfitting when the time filter was active.



(a) Weights for each microphone

(b) Microphones are plotted as vertices. Each edge between the vertices shows the individual weight contribution in thickness and color.

Figure 4.3: Selection of microphones based on contribution using ℓ_1 regularization.

In contrast to the frequency and time bins, the analysis for the individual microphone involvement requires an additional step. Because the integration of microphones in the input tensor is already done in a combinatorial fashion by design, more weights are introduced to selectively test the involvement of microphones rather than the pairs. This means, to the expected 1540 weights another 56 weights are introduced, to promote edges that connect microphone vertices of a common set and penalizes edges of uncommon vertex connections. Each weight from the set of 1540 weights is therefore additionally multiplied with the weights

of the set of 56 that correspond to the vertices in the particular combination. For example, the combination of microphones (2,3) would generate the sequence $w_{56}(2) \cdot w_{56}(3) \cdot w_{1540}(2,3)$. Figure 4.3a shows the resulting weights per microphone, ruling out 80% of the microphones and reducing the range to 11 distinct microphones, namely (1, 4, 5, 14, 15, 16, 20, 21, 25, 33, 36). Additionally, Figure 4.3b shows the position of the microphones in the array configuration and the resulting weights of the connecting edges on a color scale. The priority of common connections of the resulting set can be clearly seen.

4.3. UW-SVM (SRP-PHAT Features) versus SW-CNN (Learned Features)

The configuration of the features is the major distinction between the suggested approaches. While the UW-SVM method uses the unedited algorithm proposed by [10] for the feature formation, the SW-CNN method injects selective filters before merging the steered spectrum. Based on this similarity it is possible to compare the two based on separability. To do that, 80% of the static data is used to train a network generating the kernel parameters. The network is then applied to the remaining 20% of the data and the intermediate feature of size B is extracted (see Figure 3.5 for reference). Using the same parameter set, the basic SRP-PHAT procedure is applied to the same 20% partition of the data. As separability measures, two methods are used, showing overall separability and a class by class measure. The overall measure is the J_3 criterion of the class scatter matrices of the set of features. The J_3 criteria is defined as

$$J_3 = trace\{\Sigma_w^{-1}(\Sigma_w + \Sigma_b)\} \tag{4.3}$$

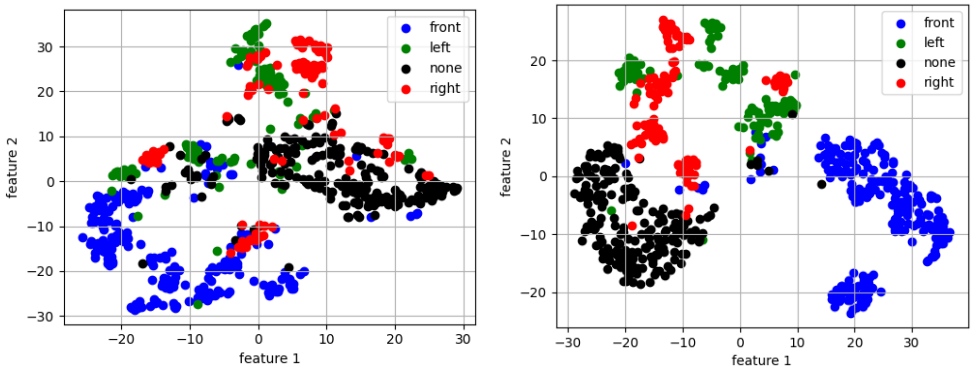
Where Σ_w is the within-class scatter matrix, which is the sum of the per class covariance. Σ_b is the between-class scatter matrix, which is the sum of the per class product of the distance between the per class mean and global mean vector in the feature space [49]. This criterion gives an overall indication of the separability of the classes. High values indicate large in between class and low within-class scattering and consequently higher separability. For two class problems, this criterion take a special form and is commonly known as Fisher's discriminant ratio, which is defined as

$$FDR = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^2}{\boldsymbol{\sigma}_1^2 + \boldsymbol{\sigma}_2^2} \tag{4.4}$$

With $\boldsymbol{\mu}_i$ being the class mean and $\boldsymbol{\sigma}_i$ the variance of class i. Table 4.2 shows the results of the test. It is worth mentioning that separability studies cannot give any definitive answers about the absolute difficulty of a classification problem, however, they are valuable in a comparative context. Therefore, the ratio between the two feature types is of most interest. The results show clearly, that the original SRP feature especially lacks behind in the separability of `left` and `right`, which is intuitively the hardest to differentiate. The learned feature shows the biggest increase between those classes of about a factor of 2.7 next to `front` and `none` with a factor of 2.9, which is arguably the most differentiable anyway. The overall measure also shows a general increase of a factor of 3.5.

Feature	`front` / `left`	`front` / `none`	`front` / `right`	`left` / `none`	`left` / `right`	`none` / `right`	J_3 criterion
SRP-PHAT	25.93	33.59	26.56	28.43	**3.97**	21.04	1379.06
Learned	34.96	98.14	39.61	40.20	**10.70**	22.75	**4827.90**

Table 4.2: Comparison of the SRP-PHAT feature vector and intermediate CNN layer in terms of discriminative power. The inter-class values refer to the sum of the class versus class Fisher ratios. The last column shows the J_3 criteria composed from the within and in between scatter matrices



(a) Feature distribution of SRP-PHAT

(b) Feature distribution of the intermediate CNN layer

Figure 4.4: Comparison of the t-SNE reduced feature spaces of the pure signal processing (4.4a) and learned (4.4b) feature formation.

Furthermore, similar observations can be made by plotting the features in the t-Distributed Stochastic Neighbor Embedding (t-SNE) [50] reduced feature space in Figure 4.4. The distribution of classes of the learned features shows a much more clustered and better defined separation than the SRP-PHAT features.

4.4. Peformance Evaluation on Static and Dynamic Data

Following that, the techniques' overall performance is evaluated using static and dynamic samples. Both methods are cross-validated on 5 folds using the specified sample set (SAD or DAB) in Table 4.3. When training the SW-CNN approach, the training set in each fold was further divided into a random 15% validation split that serves as an early stopping condition, watching the validation accuracy. The maximum number of training epochs was selected to be 70 for the static data and 100 for the dynamic data. Because no benchmarks for the scenario of the NLOS vehicle localization exist, three additional inference methods are reported alongside. First, a visual approach will be fed with the corresponding image frames that would cover the coincidental one second audio sample. A correct classification occurs if any of those image frames contain a car detection. This visual baseline is a Faster R-CNN R50-C4 model trained on the COCO dataset [51]. For the inference in this experiment, the output is further limited using a score threshold of 75% and to suppress false positives from the backdrop, all bounding boxes are filtered by a minimum height threshold of 100 pixels. The second reference method is a naive utilization of the signal processing approach. The input is the generic SRP-PHAT DOA vector, extracted from the sample. No augmentation besides the frequency filter from the UW-SVM method is used. The inference is then simply done by filtering the corresponding azimuth angle relating to the maximum value of the feature vector into the three classes `left`, `front` and `right` as shown in Figure 4.5. The `none` class was left out for this naive approach as it was not defined.

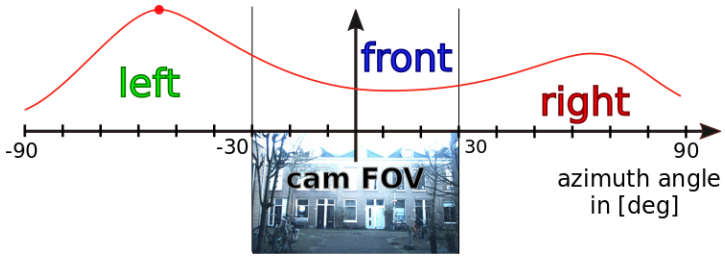


Figure 4.5: Illustration of the naive approach from Table 4.3. The red line shows an example DOA vector. Th maximum value determines the classification by association of its position on the azimuth to `left` or `right` (left or right of the camera FOV) or `front` (inside the camera FOV).

Lastly, a slightly modified version of the SELD-net [21] was evaluated on the dataset. The SELD-net method is a recurrent convolutional neural net approach, that solves the multi model problem of sound event detection and localization. An illustration of the network is shown in Figure 2.1a of Section 2.1.2. Input to the network is the same data that is available to the SW-CNN method including the same active microphones, frequency and time sequence bins. However, contrary to the SW-CNN method, the SELD-net approach directly takes the phase and magnitude components of the STFT transformed signals. The modifications are

limited to shaping the output of the localization branch to output the desired class structure instead of coordinates. The event detection branch is omitted and not included. First, the SELD-net outputs coordinates per sequence instance. To squash all sequences of the samples into one, averaging is introduced between the recurrent and fully connected block of the network. Also, the fully connected layers of the SELD-net approach are replaced with the fully connected tail of the proposed SW-CNN method. The authors propose to use a large quantity of different parameters for the network. All combinations are proven to perform very well as shown in Figure 3 of [21]. For the modified network, generally smaller parameters are used to limit the size of the parameter space. For the number of convolutional layers 32, the number of recurrent layers 64, the CNN maximum pooling size 2 for all and a dropout rate of 0.3. This compares to the original parameters to (64, 128, [8, 8, 2] and 0).

Table 4.3 shows the results of the two methods and references in a direct comparison. Strongest in the results of detecting a vehicle in camera view or no vehicle at all is, of course, the visual method. Because this system is not capable of detecting anything outside its camera FOV, no instances of `left` and `right` are classified correctly. The naive approach also does fairly well on the `front` class, however less than mediocre on distinguishing `left` and `right` samples. This is due to the fact that without a line-of-sight the soundfield is just a superposition of dispersed sound waves of reflected and diffracted parts, as previously stated. As a result, the dominant DOA, inferred from the simple DOA vector, does not directly coincide with the direction where an occluded vehicle is coming from. Both, the UW-SVM and SW-CNN approach perform much better across the board indicated by the overall accuracies. However, on both datasets, the UW-SVM technique with designed features outperforms the SW-CNN method considerably. In the dynamic data specifically, the `left` and `right` accuracies suffer the most from the change to a dynamic environment. More stable, but still less accurate shows the `none` class in both examples. Since the vehicle is now in motion, there is more overlay from sound that is generated by the mounted vehicle. The methods have to cope with this ego noise on top of the changing acoustic environment, which produces a much noisier baseline and therefore the `none` class appears less distinguishable.

Method	Data	Accuracy	$IoU_{\texttt{left}}$	$IoU_{\texttt{front}}$	$IoU_{\texttt{right}}$	$IoU_{\texttt{none}}$
UW-SVM	SAB	0.92	0.79	0.89	0.87	0.83
	DAB	0.76	0.41	0.80	0.44	0.65
SW-CNN	SAB	0.86	0.62	0.87	0.62	0.82
	DAB	0.70	0.18	0.82	0.23	0.59
DoA-only [52, 53]	SAB	0.64	0.11	0.83	0.28	-
SELD-net	SAB	0.70	0.21	0.83	0.41	0.67
Faster R-CNN [51]	SAB	0.60	0.00	0.99	0.00	0.98

Table 4.3: Comparison of the different methods with a 5 fold cross correlation.

4.5. Generalization Across Acoustic Environments

The techniques are compared in the following experiment based on how effectively they generalize across the specified environment types A and B. The dataset is therefore split into training and testing sets according to the environment association. The divisions established are shown in Table 4.4. First, both methods are trained on B type environments and tested on A, then trained on A type environments and tested on B. The same was done for static splits, as well as for the dynamic data, but separately. Furthermore, the SW-CNN approach was not split into a validation set as the training samples have already shrunken significantly in these splits. Instead, the network was trained for a fixed 50 epochs on the static training set and 80 epochs on the dynamic data without an early stopping argument.

Table 4.4 shows mediocre results across the board. While the SW-CNN method seems to be more consistent in distinguishing the `front` from the `none` class in the static data case, it struggles to find a balance in the classification between `left` and `right`. Whether training on SB and testing on SA or training on SA and testing on SB, the network converged towards favoring one over the other. On the contrary, the UW-SVM method, while not performing well in training on SB and testing on SA, still balances `left` and `right` in both static cases. Both methods perform worse overall in its equivalent dynamic case. Interestingly, the UW-SVM method struggles more to balance `front` and `none` while training on DB and testing on DA, while the opposite is true for the SW-CNN method. To reiterate, the main difference in dynamic data versus static

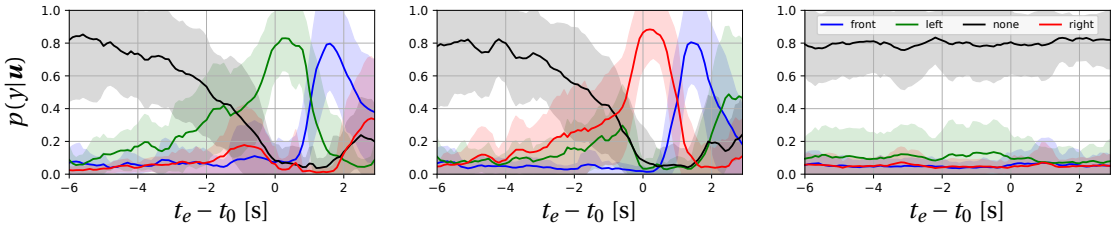
Method	Training	Test	Accuracy	$IoU_{\texttt{left}}$	$IoU_{\texttt{front}}$	$IoU_{\texttt{right}}$	$IoU_{\texttt{none}}$
UW-SVM	SB	SA	0.66	0.03	0.66	0.03	0.62
	SA	SB	0.79	0.42	0.82	0.61	0.67
	DB	DA	0.53	0.16	0.70	0.25	0.16
	DA	DB	0.56	0.21	0.50	0.29	0.46
SW-CNN	SB	SA	0.72	0.33	0.69	0.06	0.72
	SA	SB	0.76	0.17	0.83	0.47	0.74
	DB	DA	0.56	0.22	0.58	0.12	0.52
	DA	DB	0.46	0.19	0.59	0.14	0.23

Table 4.4: Generalization across environment types.

data is the moving receiver array, which can significantly impact the transmission function between source and receiver due to the changing acoustic environment. But additionally, all audio samples are overlaid with another modality of environmental noise, the addition of aerodynamic noise due to wind and the noise of the driving ego vehicle.

4.6. Generalization Across Time Horizon

As mentioned in Section 1, the goal is to identify potential dangers emerging from such circumstances as early as possible by using the advantage of detectable sound emissions emanating from blocked areas. As a result, the methods are tested on a larger time horizon of the recordings. This shall give an indication on how much time prior to a conventional system the proposed methods can confidently estimate oncoming traffic. The methods are trained on a stratified split with 328 training recordings. The remainder of 83 recordings were used as test sets for the following experiment. Because the classifiers need a second long sample, the assessment only begins at 1 second into the recording, at which time a complete audio sample has accumulated. Therefore, the 10 second long recordings are cropped to 9 seconds. Going further, the inference is applied on a sliding window with a 0.1s step size. The step size is chosen as it coincides with the camera frame rate for visualization. The confidence output of the classifiers are recorded and the average is plotted over all the test recordings in Figure 4.6. The standard deviation of the set of confidences each time step is represented by the clear colored regions in the graphs.



(a) `left` recordings, UW-SVM method (b) `right` recordings, UW-SVM method (c) `none` recordings, UW-SVM method

(d) `left` recordings, SW-CNN method (e) `right` recordings, SW-CNN method (f) `none` recordings, SW-CNN method

Figure 4.6: Shown are mean and standard deviation of predicted class probabilities at different times t_e on the test set recordings of the static data (blue is `front`, green is `left`, red is `right`, and black is `none`). Each figure shows recordings of a different situation. The approaching vehicle appears in view just after $t_e - t_0 = 0$. The top row, Figure 4.6a - 4.6c shows the results of the UW-SVM method. Figure 4.6d - 4.6f for the SW-CNN method.

Figure 4.6 shows the evolution of the classifier outputs while a car is passing by. For all occasions in the `left` and `right` recordings the confidences of the `none` class drop as early as $t = -4$s and `right` and `left` confidences slowly increase. In the period of $t = -4$s to $t = -1$s the `none` class is still dominant. However, independent of the recording type, both classes, `left` and `right` rise. This may indicate the classifier is getting aware of the changing surroundings and traffic ahead, while not being able to distinguish between the two. After $t = -1$ the confidence of the correct classes rise sharply and confirm indistinguishable presence of the target vehicle around a specific corner. The confidence of the `front` class takes over at around $t = 1$s. Although the target vehicle is visible at the very first instance after $t = t_0$, the training and testing samples have a duration of 1 second. Audio and video lines have been synchronized and aligned. Therefore, only after $t = 1$s, the audio sample does not contain any data prior to t_0 anymore. That means, seeing `left` and `right` classes past t_0 is a result of the inference appear to lag behind. Comparing the methods against each other shows a much smoother evolution for the UW-SVM approach. The confidence values of the SW-CNN approach jitter during inference, indicating much sharper decision boundaries. Overall, the `none` class progression in `left` and `right` recordings seem fairly similar in both methods, while slightly better for the SW-CNN approach in only `none` recordings. The `left` and `right` class confidences seem a bit lower and shallower for the SW-CNN approach, however higher and broader for the `front` class.

To quantify the results, the accuracies are calculated for each timestep based on the absolute classifications in this experiment. Figure 4.7 shows the absolute accuracy over time for all test recordings. Before t_0, each sample is labeled according to its recording, i.e. every timestep before t_0 in the `right` recordings is labeled as `right`, after t_0 it is labeled as `front`. In the transition period, which is indicated as the gray area in the plot, a multilabel is briefly allowed. That means between $t = t_0$ and $t = t_0 + 1.5$s of `left` and `right` recordings, both classes, `front` and the respective direction (`left` or `right`) are considered as True Positives. `none` recordings are labeled as `none` for all time steps of the entire time horizon. Both methods are compared to a visual baseline that is evaluated alongside using the same parameters as discussed in Section 4.4.

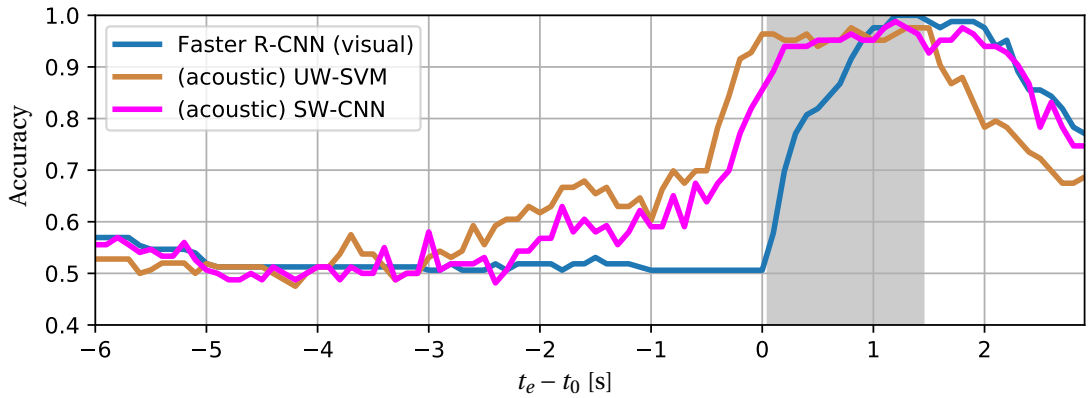
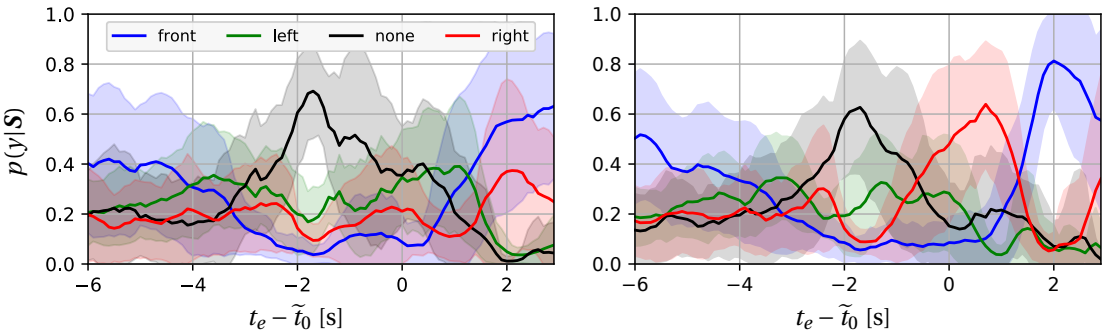


Figure 4.7: Accuracy over test time t_e of both acoustic methods and the visual baseline on 83 static recordings. The gray region indicates the other vehicle is half occluded and two labels, `front` and either `left` or `right`, are considered correct.

The visual baseline in Figure 4.7 gives an indication of the line-of-sight conditions. Right after t_0 the accuracy sharply increases. Only at $t = 1$s almost all instances result in a visual detection of the target vehicle. t_0 is labeled as the image frame at which the car is still behind the occluding wall. That means that the first frame after t_0 is only comprised of a very small portion of the target vehicle reaching beyond that corner. The Faster R-CNN does a great job in those tricky conditions. However, in addition to the occluding walls, some locations were further compromised by other occlusions such as foliage from bushes or parking cars. While not blocking appropriate labeling, this might have caused the lower rise of the visual accuracies right before $t = 1$s. Traffic conditions varied at each location and the target vehicle approached at different speeds as well. Therefore, a continuous decrease of the visual baselines towards the end of data is expected, where the target vehicle exits LOS conditions at different times. Both acoustic methods perform remarkably well, reaching high accuracies before t_0. As described earlier in Figure 4.6 the period between $t = -4$ and $t = -1$ at which the classifiers pick up a change is visible in this figure as well. In this period the accuracies increase slightly, though not reaching conclusive values. As Figure 4.6 suggests, multiple occasions occur in this period where the classifier detects `left` or `right` without strong preference for each in either recordings (`left`

or `right`). After $t = -1$s both methods sharply increase, reaching an accuracy value of 0.9 before (for the UW-SVM approach) and around $t_e = t_0$ (the SW-CNN approach). This illustrates how powerful the proposed acoustic detection pipeline is over a sole visual detection setup. The acoustic methods detect the target vehicle around one second before the visual baseline showing matching accuracies. Furthermore, the SW-CNN method keeps a higher accuracy for a longer time over the UW-SVM method for the `front` period. The accuracy line of the SW-CNN approach follows the steady decline of the visual baseline, while the UW-SVM method collapses earlier.



(a) `left` recordings (b) `right` recordings

Figure 4.8: Mean and std. dev. of predicted class probabilities at different times t_e on `left` and `right` test set recordings of the Dynamic data. The ego-vehicle reached the location of training data when $t_e - \widetilde{t}_0 = 0.5$s.

Lastly, the same analysis that was conducted for the static case was applied for the dynamic data. However, no visual baseline could be extracted. As the ego vehicle was driving towards the intersection, most of the time the FOV of the camera reached the occlusion boundaries (building edges) before the target vehicle came into LOS. When reaching the camera FOV, the target vehicle was already so close that no meaningful detection could be extracted. Also, only the better performing method is shown for the dynamic dataset, namely the UW-SVM method, because the results showed little additional insight. Figure 4.8 shows therefore the confidence progression of the UW-SVM classifier. The classifier was trained on a stratified training set of 233 recordings and tested on 59. It can be clearly seen that the classifier struggles over the entire time horizon. Especially in the beginning, the classifier shows low confidences across the board, where the `none` class should be dominant for a while. This could be partly explained due to the fact that the ego vehicle is not in motion prior to the recording, but needed to accelerate, generating more noise than during constant travel. However, although it may coincide, it is worth mentioning again, that the recordings are aligned to $\widetilde{t}_0$ and hence the time of acceleration varies by recording and of course location. In later time steps closer to $\widetilde{t}_0$ the confidence of the correct classes rises, however not towards an indistinguishable level. After passing $\widetilde{t}_0$ the `front` class correctly comes on top in both recording types similar as in the static case.

5
Conclusion

The results of the performed experiments show that the UW-SVM method is superior to the SW-CNN method. The main difference is the learning of the contributions of the S tensor that result in a directivity vector. While it would be expected to achieve higher performance and generalization capabilities with this addition, the contrary is true and the inclusion in the learning procedure does not appear to help. Both however outperform a naive acoustic classifier that processes the DOA estimates simply by maximizing the output, as conventionally done in LOS applications. Similarly, the visual reference pipeline breaks down when tasked with the early detection of occluded vehicles. Therefore, it can be concluded that Table 4.3 proves the possibility of early vehicle localization in shadow zones of occlusions using only acoustic cues. Despite succeeding only with significantly positive results under certain conditions, this thesis validates the proposed concept and provides a solid foundation for future exploration.

5.1. Discussion

First, the hyperparameters of the feature formation are addressed. Despite elaborating the choices in Section 4.2 individually per method, reducing the data size for the SW-CNN method created an imbalance between the two methods and one could argue that one performs better for the reason of having access to more data. However, that is not likely. The generic SRP-PHAT method of Dibiase [10] describes the method as the steered response power of the filter and sum beamformer, using the inverse Fourier Transform, integrating over the frequencies ω. Hence, the DOA feature is highly sensitive to the selection of the frequency range. Considering a binaural system and using the STFT with a window size equivalent to the sample size, the microphone and time dimension size would resolve to 1 for each as in Equation 3.4. Lowering the window size and increasing the microphone count simply increases the spatial and temporal resolution. The author of [10] addresses the impact of which himself in Section 6.6 in their work, elaborating that these two dimensions only robustify the outcome. Therefore, despite reducing the size of those two dimensions for the SW-CNN method, the sensitive frequency dimension is left with a larger range ($[0, 1900]$Hz) than the UW-SVM method ($[50, 1500]$Hz in comparison). So one could argue that SW-CNN actually sees more relevant data than the other.

How accurately can a purely acoustic localization system predict oncoming traffic in the specified scenario?

In Section 4.4 and Tables 4.3, it is shown that an overall accuracy of 0.92 can be achieved for the static dataset using the UW-SVM method. The per class accuracies average around 0.85 with a standard deviation of 0.04 and the highest score of 0.89 for the `front` class in this particular experiment. Reaching the highest score with the `front` class is fairly intuitive as the mode is described by having a direct line-of-sight towards the receiver, besides the presumable highest sound pressure level. It can be argued however that the `front` class has the largest variance in the feature space. Because the labeling process was aligned with a static offset from t_0, target vehicles are not expected to be at the same position in every sample as is the case for the `left` and `right` samples. Therefore, despite having a direct line-of-sight, the angle of maximum excitation spreads across the range of the field of view in the feature DOA. The `none` class performs second worst across this board. As no effort was taken to deal with, or attenuate direct, external noise effects, the classifier was tasked directly to distinguish any output corrupted by noise from any of the positive classes (`left`,

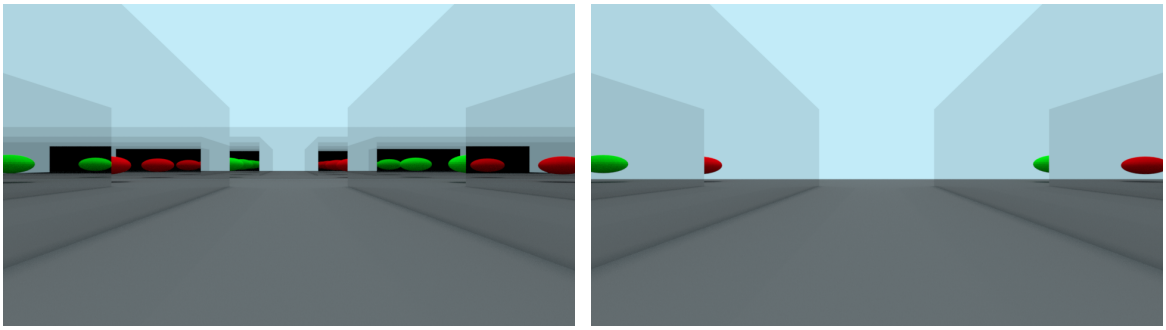
front, right). This is also true for all the methods that are part of this investigation. Despite having to deal with this disadvantage, the method actually performs exceptionally well in this inference scenario, especially considering the scores of the left and right class. While the difference between the front and none class may seem obvious on the surface, distinguishing either left or right from scattered noise of none samples is considerably much harder. As it can be seen in other experiments, the toughest job is the distinction of left and right samples. However, as Table 4.3 shows, the UW-SVM performs equally well in these classes, although the average per class accuracies are lower compared to front and none, which is 0.83 for left and right and 0.86 for front and none.

How do data-driven methods compare to classical signal processing methods for acoustic localization in non-line-of-sight situations?

Table 4.3 also features the results of a naive classification pipeline that draws the DOA angle from the maximum value of the feature vector. The range based inference breaks the classes down to either out-of (left or right) or inside-of (front) the cameras FOV, which coincides with the alley opening in the camera frame. This naive approach comes closest to a purely signal processing technique for the proposed NLOS task. This naive approach is only able to reach an accuracy of 0.64 overall, only considering the three class system without the none class. The front class stands out with a per class accuracy of 0.83, while the left and right class only reach 0.11 for left and 0.28 for right respectively. Looking back at the comments made in the previous paragraph, the high accuracy for the front class seems expected. As the target vehicle is in line-of-sight, the receiver experiences a dominant DOA from the sound source, which is also present in the camera FOV, resulting in a majority of True Positives for this mode. The performance of the left and right class however suffers from this generic inference scheme. This means a dominant DOA that is received left, outside the camera field of view does not determine a vehicle coming from the left shadow zone indistinguishable. The intersection over union used as the per class accuracy metric penalizes False Negatives. Therefore, only a small portion of left and right samples are expected to be predicted as the front class due to its high score. This means the majority of misclassifications happen between the classes left and right exclusively. This leads to the conclusion that there is no direct relation of dominant DOA to the direction, a target vehicle is coming from in the NLOS condition. Hence, signal processing techniques are not suited for the NLOS localization task without additional knowledge about the scene. The proposed learning based techniques outperform the naive signal processing one by far.

How does a different location or ego-motion of the sensor array influence the accuracy of the approach?

Section 4.5 compares the proposed methods in terms of how well either generalize across the different environment types defined in Section 3.2.3. For that, each method was trained solely on one type and tested on the other. Table 4.4 illustrates the results of the experiment. It can be clearly seen that both methods struggle to achieve similar results to those including all environment types in the data. The average overall accuracy in the static case compares slightly better for the SW-CNN method (0.74) than for the UW-SVM approach (0.73), however, that changes when the dynamic tests are included as well (0.64 for UW-SVM and 0.63 for SW-CNN). Especially notable is the decrease of the per class accuracies of the left and right class, indicating clearly that the directivity characteristic of the features gets lost if not included in the training set. This leads to the conclusion that both methods cannot generalize well across the different environment types. What can be said however for the different environment types is that in all cases it seems to be favorable to train on A types rather than on B. This is reflected by all overall accuracy results and also the per class accuracies of the left and right classes. The only exception is in the dynamic case of the SW-CNN method. This is despite having a large sample size disadvantage. Recalling Table 3.1 from Section 3.2.3, the data was recorded at two different type A locations and three different type B locations, rendering the sample sizes roughly to a ratio of 2:3. Type A locations differ due to the additional acoustic surface across the ego vehicle. This reflective surface adds to the cumulative paths that the soundwave can take, which forms the DOA vector. An illustration of how this effect unfolds can be seen in Figure 5.1. 5.1a shows a lot more mirror objects of the target blobs in the 3D render than 5.1b does. While still incorporating the first order reflections from the left and right, type A locations receive additional excitation at lower angles due to the reflections of the facing wall. Training on type B locations, missing the additional excitations can decrease the generalization for type A locations while, vice versa, the reflection pattern, that exist in both, helps to generalize much better while training on type A locations.



(a) 3D render of the type A environment

(b) 3D render of the type B environment

Figure 5.1: The figures show rendered 3D environments in which the wall surfaces have 100% specular reflection properties. The layout resemble the two environments A (5.1a) and B (5.1b). The target vehicles are symbolized by elongated spheroids. The color scheme resembles previous figures. The left spheroid is colored green and the right spheroid red.

Table 4.3 shows the difference of the method performance when applied to the dynamic versus the static dataset. Overall the accuracy decreases by 0.16 for both of the methods in the dynamic case. The effects on the per class accuracies are also comparable for both methods. While the `front` class accuracy seems to be the best performing class in the dynamic case, the `none` class suffers in accuracy points and the `left` and `right` class suffers the most. As previously stated, the position of the receiver, as well as the position of the target sound source, can greatly affect the transfer function between them as the reflection and diffraction pattern change over time. However, it can still be expected that there is a pattern that would make a classification possible as the underlying principles do not differ between the static and the dynamic case. Therefore, the compromising factor for the decrease in accuracy may lie in external factors. First, the receiver suffers now from additional noise as the ego vehicle travels as well. The noise sources are ego tyre noise, ego engine noise and notably the aerodynamic noise at the microphones themselves, as they are now subject to a steady relative flow. This is additionally amplified by the narrow street canyon, aggregating the generated noise that is picked up by the microphone array. Furthermore, the dynamic dataset is recorded and labeled by different means. Starting with the difficulty of coordination of two vehicles. `left` and `right` samples in the dynamic dataset may come with a variance of the target vehicle position, being more or less 2 seconds early or late due to these difficulties. Despite having an odometer in the ego vehicle, the target vehicle was not tracked over the individual recordings. Therefore, an accurate estimate of the spread is not available. That means, the samples at the alignment time $\widetilde{t}_0$ may show different reflection and diffraction patterns in the data, which could corrupt the feature space and hence inference accuracy. This is however very crucial as the hazard in the dynamic scenario is only present if the temporal paths of ego and target vehicle intersect. Therefore, it remains to be seen if the system may perform better if the vehicles are better aligned.

How much earlier can the proposed method confidently detect traffic with respect to a visual method?
The potential of early vehicle identification in dangerous situations is the most significant improvement that the findings of this thesis are anticipated to bring to autonomous or aided driving. As a reference, the results are compared to a state of the art visual detection system, representing the conventional line-of-sight sensor systems. Table 4.3 shows the direct comparison of the acoustic versus the visual methods. As expected, the visual system shows almost flawless accuracy for the `front` and `none` class. However, as there is no means of classifying vehicles in NLOS areas, the `left` and `right` classes are 0, rendering the overall accuracy to a bare 0.6. The important question however is, how much earlier the acoustic method reach as confident results about an approaching vehicle as the visual system does in line-of-sight conditions.

Section 4.6 looks at the performance of the classifier at different time steps of the recordings. Figure 4.6a to 4.6c show the per class confidence averages over all test recordings of the UW-SVM method and 4.6d to 4.6f the same for the SW-CNN method, both in the static case. The confidence output shows intuitively good results across the entire time horizon as previously discussed. While it is expected that the classifier perform best at the exact time frame that they have been trained on, i.e. $t = t_0$ for the `left` and `right` classes and $t = t_0 + 1.5$s for `front`, the progression of the output paints a little different picture. When approaching these distinct time steps, the confidence values sharply increase, however, remain high for a longer while.

For example, at t_0 of the `left` and `right` recordings, the `left` and `right` confidences remain high, despite the vehicle already being in line-of-sight. This is due to the sample length and the causal structure of the data. Recalling the data setup, the audio includes the entire second before the visual image frame. That means, despite the vehicle already showing in camera view, a large portion of the audio sample consists of NLOS data. This fact inspired the labeling procedure for the evaluation in Figure 4.7 in the transition period. Figure 4.7 shows the accuracies per time step in a comparison of the two proposed methods versus the visual method. In contrast to the confidence plots, the absolute classifications are used to give insight into the actual time difference between them. It shows that the UW-SVM method can reach an accuracy value across all recordings of 0.9 about one second, the SW-CNN method about 0.65 seconds earlier than the visual method. Furthermore, while the visual method accuracy can only increase after t_0, both acoustic methods ramp already up at around $t = -4$s, showing an even earlier potential for detection. For the dynamic situation unfortunately no such statements are possible at this point.

Is an end-to-end approach viable without any preprocessing of the data for the task at hand?

Comparing the UW-SVM and SW-CNN methods directly, it can be undoubtedly stated that the UW-SVM performs better across the board, generalizing better across different environments and shows an earlier detection rate. However, the features had to be carefully engineered in order to acquire the expected results. It remains to be seen how much further this method will be the more viable if for example more variance is introduced to the data. The situation of highly reverberant environments has been challenging for acoustic localization techniques for years. Many researchers have converged to employ deep learning techniques to cope with these complex environments as described in Section 2.1.2. Using data-driven methods is crucial to deal with these highly non-linear phenomena of sound propagation. That is especially true for NLOS conditions, where reverberation is not only suppressed but harvested to tell the classes apart. As models get more complex, the requirements on the data rise. Comparing the dataset to other acoustic data corpi, the one recorded in this thesis can be considered fairly small with 411 original recordings or 623 samples. Therefore, the modified SELD-net in Table 4.3 shows such poor results operating directly on the STFT spectrogram input. Nevertheless, Section 4.3 shows the difference a learning based approach can make. The standalone SRP-PHAT feature is compared to the trained counterpart, which is the intermediate layer of the SW-CNN network. As a measure of descriptiveness, the Fisher ratio is used. Table 4.2 shows the results as an overall and per class numeric measure. The improvement between the `left` and `right` class shows clearly that the network is capable of extracting more meaningful information from the data by mere weighting of the input tensor. The difficult classes, `left` and `right`, are spread further apart as shown in the t-SNE plots in Figure 4.4 for the learned feature space than they do in the generic SRP-PHAT method. However, the SW-CNN method is unfavorable as the input size grows massively compared to the actual audio input data. Despite the educated reduction of features (see Section 4.2), the input to the network renders to 4.5 million data points, compared to the original audio data points of 0.5 million (56 microphones at one second sample and 48000 sample rate) or the reduced STFT spectrogram for the SELD-net only 2,646 data points. Employing an end to end network could therefore greatly benefit in reducing the computational complexity of the feature extraction, as well as inference. However, the complexity of the problem still requires the assistence of the generic signal processing pipeline to produce meaningful results in the absence of a large enough data corpus.

5.2. Future Work

As previously stated, the goal remains to find a network that operates in an end to end fashion, either using the raw audio signals as inputs or the sequence spectrograms. This would eliminate the angular sampling procedure as well as the binomial expansion of microphone pairs, which results in generally costly preprocessing and clutter in the input stream. However, as shown in Section 4.4 more complex models using less processed data have shown to be less viable due to the limited amount of data. Furthermore, the experiments have shown that both dynamic data, as well as different environments, can decrease the performance of the methods significantly. The solution may be to simply increase dataset size and variance.

The cheapest way to increase data size would be to automate the recording and labeling procedure on the research vehicle so that a simple drive through selected locations would be sufficient to add to the corpus. This would increase the variance in all modalities: target vehicle type, environments and external noise, ego vehicle dynamic state and much more. With the aid of the additional camera and the other internal sensors of the research vehicle, such as the GPS, a reliable implementation can be found.

In order to increase the generalization performance of the proposed methods, besides more representative data, a few other methods could be employed to support the classifier. It has been established that the underlying principle for detecting and localizing occluded vehicles is based on the reflection and diffraction pattern of the acoustic scene. Changing the acoustic scene geometrically changes the meta of the acoustic path and can heavily alter what is picked up by the sensor array. As acoustic scenes in real world environments tend to be very complex, so would be a physical model representation or transfer function. Manifold learning is a strategy that tries to lump highly complex parameter spaces into a lower dimensional space. This is especially useful if the parameter space only has a few degrees of freedom. For example, the transfer function of an acoustic scene can be learned, that lies on the approximate manifold A, which is characterized by the layout of the environment A versus B. These parameters could further guide the network in changing environments, that differ in auxiliary factors but still keep the general shape of the scene. Furthermore, this can be extended to a semi-supervised learning scheme as the manifold representation does not directly require hard labels in contrast to the inference of the classes. The idea to support acoustic localization methods by the acoustic transfer function is not a new one and has succeeded in several occasions, e.g. [38]. Most attempts try to deal with line-of-sight situations in difficult indoor locations, therefore it is not certain if these apply directly to NLOS situations.

If this method would prove to provide additional support, this could benefit not only the generalization across different locations but also for dynamic situations. The biggest drawback of the dynamic data analysis is however the practicality of the recordings as described in Section 5.1. Much more valuable seems to be if static recordings would be transferable to dynamic data as well. Besides obvious techniques of noise reduction, one could think about means of domain adaptation. Using the recorded data corpus, one could find structural similarity between the static and dynamic recordings of the same class and locations and find an alternative, domain-invariant feature representation instead. This way, a minimal amount of difficult to record dynamic data would suffice with a larger corpus of easily recordable static data. While finding a domain-invariant feature may be challenging, there are other, more advanced methods for domain adaptations to explore, such as adversarial or reconstructive domain adaptation that may prove more beneficial for the task.

Moving on from the constraint situation of T-crossings, there are plenty of applications in which acoustic localization can enhance the receptive capabilities in NLOS conditions. Most notably are salient sound events that can occur in traffic situations. Emergency vehicles use sirens to make participants aware of their presence. These sirens can be heard across multiple city blocks, however, making out from which direction the siren is coming from can prove challenging at times. Acoustic localization techniques could help identify directions early to launch appropriate evasive actions in a timely manner. As sirens are designed to be heard from far away, different factors may be relevant in developing a localization method. A larger scale could pronounce malicious effects, for example, wind that alters the acoustic path. This situation is also not constraint to the low speed inner city condition that has been the subject for this thesis, therefore more thoughts have to be put on noise characteristics and suppression. Other notable scenarios at which acoustic localization may be beneficial as an NLOS sensor would be behind accidents. Often, a small incident can lead to a much larger one because other road users fail to recognise an extraordinary situation. Because it is occluded by other road users, they may run mindlessly into it. Screeching tires or chassis impact can again serve as early indicators of hazardous situations in proximity.

Bibliography

[1] J. Guo, U. Kurup, and M. Shah, "Is it Safe to Drive? An Overview of Factors, Metrics, and Datasets for Driveability Assessment in Autonomous Driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3135–3151, 2020.

[2] J. Van Brummelen, M. O'Brien, D. Gruyer, and H. Najjaran, "Autonomous vehicle perception: The technology of today and tomorrow," *Transportation Research Part C: Emerging Technologies*, vol. 89, no. 0968-090X, pp. 384–406, 2018.

[3] A. Wang, Y. Sun, A. Kortylewski, and A. Yuille, "Robust object detection under occlusion with context-aware compositionalnets," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 12 642–12 651, 2020.

[4] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] E. A. King, A. Tatoglu, D. Iglesias, and A. Matriss, "Audio-visual based non-line-of-sight sound source localization: A feasibility study," *Applied Acoustics*, vol. 171, p. 107674, 2021.

[6] D. P. Hewett, "Sound propagation in an urban environment," Ph.D. dissertation, University of Oxford, 2010.

[7] Z.-q. Wang and D. Wang, "Learning Based Blind Speaker Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457–468, 2019.

[8] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound Source Localization in a Multipath Environment Using Convolutional Neural Networks," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 2386–2390, 2018.

[9] M. Helfer, "General Aspects of Vehicle Aeroacoustics," Sint-Genesius-Rode, Belgium: von Karman Institute for Fluid Dynamics, Tech. Rep., 2005.

[10] J. H. Dibiase, "A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays," Ph.D. dissertation, Brown University, Providence, Rhode Island, 2000.

[11] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[12] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transaction on Antennas and Propagation*, vol. AP-34, no. 3, pp. 190–194, 1986.

[13] R. Roy and T. Kailath, "ESPRIT-Estimation of signal parameters via rotational invariance techniques," *Adaptive Antennas for Wireless Communications*, vol. 37, no. 7, pp. 224–235, 2009.

[14] Y. S. Yoon, L. M. Kaplan, and J. H. McClellan, "TOPS: New DOA estimator for wideband signals," *IEEE Transactions on Signal Processing*, vol. 54, no. 6 I, pp. 1977–1989, 2006.

[15] H. Wang and M. Kaveh, "Coherent Signal-Subspace Processing for the Detection Wide-Band Sources," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 4, pp. 823–831, 1985.

[16] C. E. Chen, F. Lorenzelli, R. E. Hudson, and K. Yao, "Maximum Likelihood DOA Estimation of Multiple Wideband Sources in the Presence of Nonuniform Sensor Noise," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, 2007.

[17] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 392–402, 2014.

[18] S. Argentieri, P. Danès, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech and Language*, vol. 34, no. 1, pp. 87–112, 2015.

[19] C. Evers, H. W. Lollmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA Challenge: Acoustic Source Localization and Tracking," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 28, pp. 1620–1643, 2020.

[20] *Detection And Classification of Acoustic Events*, http://dcase.community/challenge2019, 2019 (accessed June 22, 2020).

[21] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks," *IEEE Journal on Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[22] W. He, P. Motlicek, and J. M. Odobez, "Joint localization and classification of multiple sound sources using a multi-task neural network," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-Septe, pp. 312–316, 2018.

[23] W. He, P. Motlicek, and J. M. Odobez, "Deep Neural Networks for Multiple Speaker Detection and Localization," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 74–79, 2018.

[24] Y. Cao, T. Iqbal, Q. Kong, M. B. Galindo, W. Wang, and M. D. Plumbley, "Two-Stage Sound Event Localization and Detection using Intensity Vector and Generalized Cross-Correlation," *Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 1–4, 2019.

[25] M. SIMEONI, S. Kashani, P. Hurley, and M. Vetterli, "DeepWave: A Recurrent Neural-Network for Real-Time Acoustic Imaging," *NeurIPS 2019*, no. 200021, pp. 15 274–15 286, 2019.

[26] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards End-to-End Acoustic Localization Using Deep Learning: From Audio Signals to Source Position Coordinates," *Sensors*, vol. 18, no. 10, 2018.

[27] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, 2017.

[28] L. Wang and A. Cavallaro, "Acoustic sensing from a multi-rotor drone," *IEEE Sensors Journal*, vol. 18, no. 11, pp. 4570–4582, 2018.

[29] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Acoustic Source Localization From Multirotor UAVs," *IEEE Transactions on Industrial Electronics*, vol. 0046, no. c, pp. 1–1, 2019.

[30] A. K. Pandey and R. Gelin, "A Mass-Produced Sociable Humanoid Robot: Pepper: the First Machine of Its Kind," *IEEE Robotics and Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018.

[31] M. Kreković, I. Dokmanić, and M. Vetterli, "EchoSLAM: Simultaneous localization and mapping with acoustic echoes," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 11–15, 2016.

[32] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 9, pp. 1484–1498, 2018.

[33] Y. Jang, J. Kim, and J. Kim, "The development of the vehicle sound source localization system," *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2015*, pp. 1241–1244, 2016.

[34] V. Singh, K. E. Knisely, S. H. Yönak, K. Grosh, and D. R. Dowling, "Non-line-of-sight sound source localization using matched-field processing," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 292–302, 2012.

[35] L. C. Mak and T. Furukawa, "A time-of-arrival-based positioning technique with non-line-of-sight mitigation using low-frequency sound," *Advanced Robotics*, vol. 22, no. 5, pp. 507–526, 2008.

[36] I. An, D. Lee, J. W. Choi, D. Manocha, and S. E. Yoon, "Diffraction-aware sound localization for a non-line-of-sight source," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 4061–4067, 2019.

[37] N. Scheiner, F. Kraus, F. Wei, B. Phan, F. Mannan, N. Appenrodt, W. Ritter, J. DIckmann, K. DIetmayer, B. Sick, and F. Heide, "Seeing around Street Corners: Non-Line-of-Sight Detection and Tracking In-the-Wild Using Doppler Radar," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2065–2074, 2020.

[38] B. Laufer, R. Talmon, and S. Gannot, "Relative transfer function modeling for supervised source localization," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, 2013.

[39] D. I. Hanson, R. S. James, D. I. Hanson, and C. Nesmith, "Tire/Pavement Noise Study," National Center for Asphalt Technology, Auburn University, Tech. Rep., 2004.

[40] P. M. Hofman and A. J. Van Opstal, "Binaural weighting of pinna cues in human sound localization," *Experimental Brain Research*, vol. 148, no. 4, pp. 458–470, 2003.

[41] D. Muramatsu and K. Sasaki, "Transmission analysis in human body communication for head-mounted wearable devices," *Electronics (Switzerland)*, vol. 10, no. 10, 2021.

[42] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound Event Detection and Time-Frequency Segmentation from Weakly Labelled Data," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 4, pp. 777–787, 2019.

[43] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, "Machine learning in acoustics: Theory and applications," *The Journal of the Acoustical Society of America*, vol. 146, no. 5, pp. 3590–3628, 2019.

[44] U. Sandberg, W. Kropp, and K. Larsson, "The Multi-Coincidence Peak around 1000 Hz in Tyre/Road Noise Spectra," *Acta Acustica (Stuttgart)*, vol. 89, pp. 1–8, 2003.

[45] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Methods," *Advances in Large Margin Classifiers*, pp. 61–74, 1999.

[46] L. Ferranti, B. Brito, E. Pool, Y. Zheng, R. M. Ensing, R. Happee, B. Shyrokau, J. F. Kooij, J. Alonso-Mora, and D. M. Gavrila, "SafeVRU: A research platform for the interaction of self-driving vehicles with vulnerable road users," *IEEE Intelligent Vehicles Symposium, Proceedings*, pp. 1660–1666, 2019.

[47] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," *IEEE International Conference on Intelligent Robots and Systems*, pp. 4193–4198, 2016.

[48] L. Oth, P. Furgale, L. Kneip, and R. Siegwart, "Rolling shutter camera calibration," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1360–1367, 2013.

[49] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. London: Elsevier, 2008, vol. fourth Edition.

[50] L. V. D. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research 1*, pp. 267–84, 2008.

[51] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[52] M. Mizumachi, A. Kaminuma, N. Ono, and S. Ando, "Robust sensing of approaching vehicles relying on acoustic cues," *Sensors*, vol. 14, no. 6, pp. 9546–9561, 2014.

[53] Y. Jang, J. Kim, and J. Kim, "The development of the vehicle sound source localization system," in *APSIPA*. IEEE, 2015, pp. 1241–1244.

Appendix A

Publication

Part of this thesis was published in the Journal Robotics and Automation Letters (RA-L) and presented at the International Conference on Robotics and Automation (ICRA). The publication is appended in the following pages.

Y. Schulz, A. K. Mattar, T. M. Hehn and J. F. P. Kooij, "Hearing What You Cannot See: Acoustic Vehicle Detection Around Corners," in *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2587-2594, April 2021, doi: 10.1109/LRA.2021.3062254.

Hearing What You Cannot See: Acoustic Vehicle Detection Around Corners

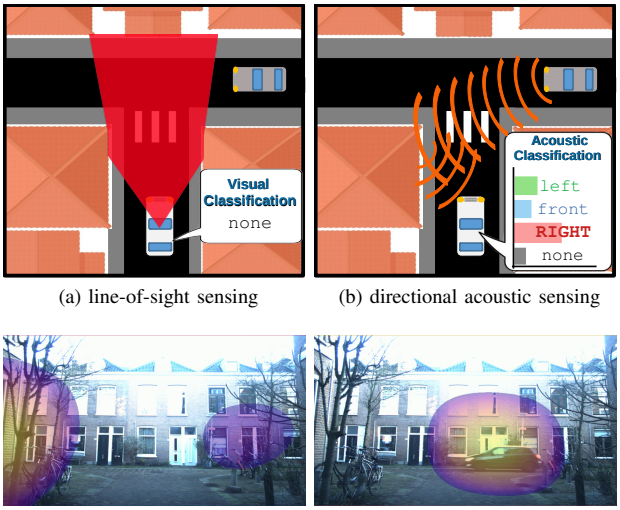
Yannick Schulz*[1], Avinash Kini Mattar*[1], Thomas M. Hehn*[1], and Julian F. P. Kooij[1]

***Abstract*— This work proposes to use passive acoustic perception as an additional sensing modality for intelligent vehicles. We demonstrate that approaching vehicles behind blind corners can be detected by sound before such vehicles enter in line-of-sight. We have equipped a research vehicle with a roof-mounted microphone array, and show on data collected with this sensor setup that wall reflections provide information on the presence and direction of occluded approaching vehicles. A novel method is presented to classify if and from what direction a vehicle is approaching before it is visible, using as input Direction-of-Arrival features that can be efficiently computed from the streaming microphone array data. Since the local geometry around the ego-vehicle affects the perceived patterns, we systematically study several environment types, and investigate generalization across these environments. With a static ego-vehicle, an accuracy of 0.92 is achieved on the hidden vehicle classification task. Compared to a state-of-the-art visual detector, Faster R-CNN, our pipeline achieves the same accuracy more than one second ahead, providing crucial reaction time for the situations we study. While the ego-vehicle is driving, we demonstrate positive results on acoustic detection, still achieving an accuracy of 0.84 within one environment type. We further study failure cases across environments to identify future research directions.**

I. INTRODUCTION

Highly automated and self-driving vehicles currently rely on three complementary main sensors to identify visible objects, namely camera, lidar, and radar. However, the capabilities of these conventional sensors can be limited in urban environments when sight is obstructed by narrow streets, trees, parked vehicles, and other traffic. Approaching road users may therefore remain undetected by the main sensors, resulting in dangerous situations and last-moment emergency maneuvers [1]. While future wireless vehicle-to-everything communication (V2X) might mitigate this problem, creating a robust omnipresent communication layer is still an open problem [2] and excludes road users without wireless capabilities. Acoustic perception does not rely on line-of-sight and provides a wide range of complementary and important cues on nearby traffic: There are salient sounds with specified meanings, e.g. sirens, car horns, and reverse driving warning beeps of trucks, but also inadvertent sounds from tire-road contact and engine use.

In this work, we propose to use multiple cheap microphones to capture sound as an auxiliary sensing modality for early detection of approaching vehicles behind blind corners in urban environments. Crucially, we show that a data-driven pattern recognition approach can successfully identify such situations from the acoustic reflection patterns on building walls and provide early warnings before conventional line-of-sight sensing is able to (see Figure 1). While a vehicle should always exit narrow streets cautiously, early warnings would reduce the risk of a last-moment emergency brake.



(a) line-of-sight sensing (b) directional acoustic sensing

(c) sound localization with a vehicle-mounted microphone array detects the wall reflection of an approaching vehicle behind a corner before it appears

Fig. 1. When an intelligent vehicle approaches a narrow urban intersection, (a) traditional line-of-sight sensors cannot detect approaching traffic due to occlusion, while (b) acoustic cues can provide early warnings. (c) Real-time beamforming reveals reflections of the acoustic signal on the walls, especially salient on the side opposing the approaching vehicle. Learning to recognize these patterns from data enables detection before line-of-sight.

II. RELATED WORKS

We here focus on passive acoustic sensing in mobile robotics [3], [4], [5] to detect and localize nearby sounds, which we distinguish from active acoustic sensing using self-generated sound signals, e.g. [6]. While mobile robotic platforms in outdoor environments may suffer from vibrations and wind, various works have demonstrated detection and localization of salient sounds on moving drones [7] and wheeled platforms [8], [9].

Although acoustic cues are known to be crucial for traffic awareness by pedestrians and cyclist [10], only few works have explored passive acoustic sensing as a sensor for Intelligent Vehicles (IVs). [9], [11], [12] focus on detection and tracking in direct line-of-sight. [13], [14] address detection behind corners from a static observer. [13] only shows experiments without directional estimation. [14] tries

*) Shared first authors. 1) Intelligent Vehicles Group, TU Delft, The Netherlands. Primary contact: J.F.P.Kooij@tudelft.nl

to accurately model wave refractions, but experiments in an artificial lab setup show limited success. Both [13], [14] rely on strong modeling assumptions, ignoring that other informative patterns could be present in the acoustic data. Acoustic traffic perception is furthermore used for road-side traffic monitoring, e.g. to count vehicles and estimate traffic density [15], [16]. While the increase in Electric Vehicles (EVs) may reduce overall traffic noise, [17] shows that at 20-30km/h the noise levels for EV and internal combustion vehicles are already similar due to tire-road contact. [18] finds that at lower speeds the difference is only about 4-5 dB, though many EVs also suffer from audible narrow peaks in the spectrum. As low speed EVs can impact acoustic awareness of humans too [10], legal minimum sound requirements for EVs are being proposed [19], [20].

Direction-of-Arrival estimation is a key task for sound source localization, and over the past decades many algorithms have been proposed [3], [21], such as the Steered-Response Power Phase Transform (SRP-PHAT) [22] which is well-suited for reverberant environments with possibly distant unknown sound sources. Still, in urban settings nearby walls, corners, and surfaces distort sound signals through reflections and diffraction [23]. Accounting for such distortions has shown to improve localization [8], [24], but only in controlled indoor environments where detailed knowledge of the surrounding geometry is available.

Recently, data-driven methods have shown promising results in challenging real-world conditions for various acoustic tasks. For instance, learned sound models assist monaural source separation [25] and source localization from direction-dependent attenuations by fixed structures [26]. Increasingly, deep learning is used for audio classification [27], [28], and localization [29] of sources in line-of-sight, in which case visual detectors can replace manual labeling [30], [31]. Analogous to our work, [32] presents a first deep learning method for sensing around corners but with automotive radar. Thus, while the effect of occlusions on sensor measurements is difficult to model [14], data-driven approaches appear to be a good alternative.

This paper provides the following contributions: First, we demonstrate in real-world outdoor conditions that a vehicle-mounted microphone array can detect the sound of approaching vehicles behind blind corners from reflections on nearby surfaces before line-of-sight detection is feasible. This is a key advantage for IVs, where passive acoustic sensing is still relatively under-explored. Our experiments investigate the impact on accuracy and detection time for various conditions, such as different acoustic environments, driving versus static ego-vehicle, and compare to current visual and acoustic baselines.

Second, we propose a data-driven detection pipeline to efficiently address this task and show that it outperforms model-driven acoustic signal processing. Unlike existing data-driven approaches, we cannot use visual detectors for positional labeling [30] or transfer learning [31], since our targets are visually occluded. Instead, we cast the task as a multi-class classification problem to identify if and from what corner a vehicle is approaching. We demonstrate that Direction-of-Arrival estimation can provide robust features to classify sound reflection patterns, even without end-to-end feature learning and large amounts of data.

Third, for our experiments we collected a new audio-visual dataset in real-world urban environments.[1] To collect data, we mounted a front-facing microphone array on our research vehicle, which additionally has a front-facing camera. This prototype setup facilitates qualitative and quantitative experimentation of different acoustic perception tasks.

III. Approach

Ideally, an ego-vehicle driving through an area with occluding structures is able to early predict *if* and from *where* another vehicle is approaching, even if it is from behind a blind corner as illustrated in Figure 1. Concretely, this work aims to distinguish three situations as early as possible using ego-vehicle sensors only:

- an occluded vehicle approaches from behind a corner on the *left*, and only moves into view last-moment when the ego-vehicle is about to reach the junction,
- same, but vehicle approaches behind a *right* corner,
- no vehicle is approaching.

We propose to consider this task an online classification problem. As the ego-vehicle approaches a blind corner, the acoustic measurements made over short time spans should be assigned to one in a set of four classes, $\mathcal{C} = \{\texttt{left}, \texttt{front}, \texttt{right}, \texttt{none}\}$, where `left`/`right` indicates a still occluded (i.e. not yet in direct line-of-sight) approaching vehicle behind a corner on the left/right, `front` that the vehicle is already in direct line-of-sight, and `none` that no vehicle is approaching.

In Section III-A we shall first consider two line-of-sight baseline approaches for detecting vehicles. Section III-B then elaborates our proposed extension to acoustic non-line-of-sight detection. Section III-C provides details of our vehicle's novel acoustic sensor setup used for data collection.

A. Line-of-sight detection

We first consider how the task would be addressed with line-of-sight vehicle detection using either conventional cameras, or using past work on acoustic vehicle detection.

a) Visual detection baseline: Cameras are currently one of the de-facto choices for detecting vehicles and other objects within line-of-sight. Data-driven Convolutional Neural Networks have proven to be highly effective on images. However, visual detection can only detect vehicles that are already (partially) visible, and thus only distinguishes between `front` and `none`. To demonstrate this, we use Faster R-CNN [33], a state-of-the-art visual object detector, on the ego-vehicle's front-facing camera as a visual baseline.

[1] Code & data: github.com/tudelft-iv/occluded_vehicle_acoustic_detection

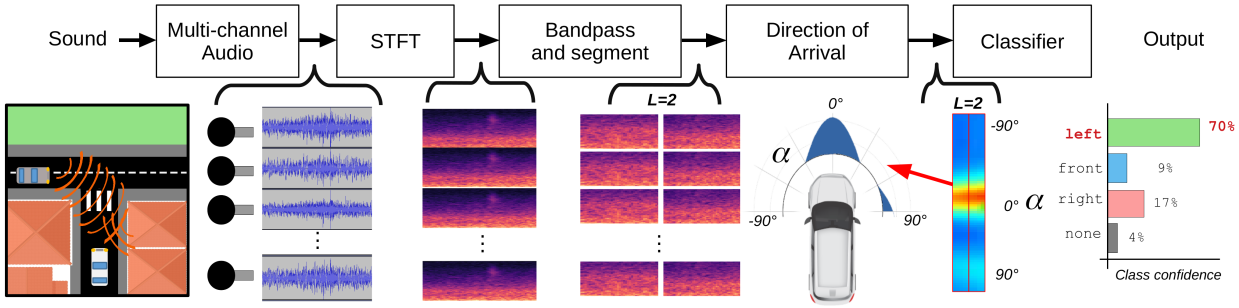


Fig. 2. Overview of our acoustic detection pipeline, see Section III-B for an explanation of the steps.

b) Acoustic detection baseline: Next, we consider that the ego-vehicle is equipped with an array of M microphones. As limited training data hinders learning features (unlike [30], [31]), we leverage beamforming to estimate the Direction-of-Arrival (DoA) of tire and engine sounds originating from the approaching vehicle. DoA estimation directly identifies the presence and direction of such sound sources, and has been shown to work robustly in unoccluded conditions [11], [9]. Since sounds can be heard around corners, and low frequencies diffract ("bend") around corners [23], one might wonder: Does the DoA of the sound of an occluded vehicle correctly identify from where the vehicle is approaching? To test this hypothesis for our target real-world application, our second baseline follows [11], [9] and directly uses the most salient DoA angle estimate.

Specifically, the implementation uses the Steered-Response Power-Phase Transform (SRP-PHAT) [22] for DoA estimation. SRP-PHAT relates the spatial layout of sets of microphone pairs and the temporal offsets of the corresponding audio signals to their relative distance to the sound source. To apply SRP-PHAT on M continuous synchronized signals, only the most recent δt seconds are processed. On each signal, a Short-Time Fourier Transform (STFT) is computed with a Hann windowing function, and a frequency bandpass for the $[f_{min}, f_{max}]$ Hz range. Using the generalized cross-correlation of the M STFTs, SRP-PHAT computes the DoA energy $r(\alpha)$ for any given azimuth angle α around the vehicle. Here $\alpha = -90°/0°/+90°$ indicates an angle towards the left/front/right of the vehicle respectively. If the hypothesis holds that the overall salient sound direction $\alpha_{\max} = \arg\max r(\alpha)$ remains intact due to diffraction, one only needs to determine if $\alpha_{\max}$ is beyond some sufficient threshold α_{th}. The baseline thus assigns class `left` if $\alpha_{\max} < -\alpha_{\text{th}}$, `front` if $-\alpha_{\text{th}} \leq \alpha_{\max} \leq +\alpha_{\text{th}}$, and `right` if $\alpha_{\max} > +\alpha_{\text{th}}$. We shall evaluate this baseline on the easier task of only separating these three classes, and ignore the `none` class.

B. Non-line-of-sight acoustic detection

We argue that in contrast to line-of-sight detection, DoA estimation alone is unsuited for occluded vehicle detection (and confirm this in Section IV-C). Salient sounds produce sound wave reflections on surfaces, such as walls (see Figure 1c), thus the DoA does not indicate the actual location of the source. Modelling the sound propagation [8] while driving through uncontrolled outdoor environments is challenging, especially as accurate models of the local geometry are missing. Therefore, we take a data-driven approach and treat the *full energy distribution* from SRP-PHAT as robust features for our classifier that capture all reflections.

An overview of the proposed processing pipeline is shown in Figure 2. We again create M STFTs, using a temporal windows of δt seconds, Hann windowing function and a frequency bandpass of $[f_{min}, f_{max}]$ Hz. Notably, we do not apply any other form of noise filtering or suppression. To capture temporal changes in the reflection pattern, we split the STFTs along the temporal dimension into L non-overlapping segments. For each segment, we compute the DoA energy at multiple azimuth angles α in front of the vehicle. The azimuth range $[-90°, +90°]$ is divided into B equal bins $\alpha_1, \cdots, \alpha_B$. From the original M signals, we thus obtain L response vectors $\boldsymbol{r}_l = [r_l(\alpha_1), \cdots, r_l(\alpha_B)]^\top$. Finally, these are concatenated to a $(L \times B)$-dimensional feature vector $\boldsymbol{x} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_L]^\top$, for which a Support Vector Machine is trained to predict $\mathcal{C}$. Note that increasing the temporal resolution by having more segments L comes at the trade-off of a increased final feature vector size and reduced DoA estimation quality due to shorter time windows.

C. Acoustic perception research vehicle

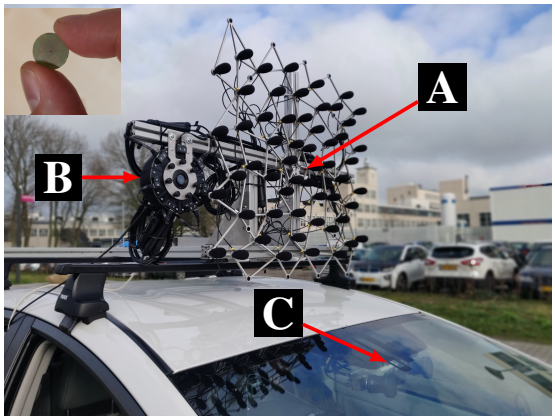


Fig. 3. Sensor setup of our test vehicle. A: Center of the 56 MEMS acoustic array. B: signal processing unit. C: front camera behind windscreen. Inset: the diameter of a single MEMS microphone is only 12mm.

To collect real-world data and demonstrate non-line-of-sight detection, a custom microphone array was mounted on the roof rack of our research vehicle [34], a hybrid electric Toyota Prius. The microphone array hardware consists of 56 ADMP441 MEMS microphones, supports data acquisition at 48 kHz sample rate, 24 bits resolution, and synchronous sampling. It was bought from *CAE Software & Systems GmbH* with a metal frame. On this $0.8m \times 0.7m$ frame the microphones are distributed semi-randomly while the microphone density remains homogeneous. The general purpose layout was designed by the company through stochastic optimization to have large variance in inter-microphone distances and serve a wide range of acoustic imaging tasks. The vehicle is also equipped with a front-facing camera for data collection and processing. The center of the microphone array is about 1.78m above the ground, and 0.54m above and 0.50m behind the used front camera, see Figure 3. As depicted in the Figure's inset, the microphones themselves are only 12mm wide. They cost about US$1 each.

A signal processing unit receives the analog microphone signals, and sends the data over Ethernet to a PC running the Robot Operating System (ROS). Using ROS, the synchronized microphone signals are collected together with other vehicle sensor data. Processing is done in python, using *pyroomacoustics* [21] for acoustic feature extraction, and *scikit-learn* [35] for classifier training.

We emphasize that this setup is not intended as a production prototype, but provides research benefits: The 2D planar arrangement provides both horizontal and vertical high-resolution DoA responses, which can be overlaid as 2D heatmaps [36] on the front camera image to visually study the salient sources (Section IV-A). By testing subsets of microphones, we can assess the impact of the number of microphones and their relative placement (Section IV-G). In the future, the array should only use a few microphones at various locations around the vehicle.

IV. Experiments

To validate our method, we created a novel dataset with our acoustic research vehicle in real-world urban environments. We first illustrate the quality of acoustic beamforming in such conditions before turning to our main experiments.

A. Line-of-sight localization – qualitative results

As explained in Section III-C, the heatmaps of the 2D DoA results can be overlaid with the camera images. Figure 4 shows some interesting qualitative findings in real urban conditions. The examples highlight that beamforming can indeed pick up various important acoustic events for autonomous driving in line-of-sight, such as the presence of vehicles and some vulnerable road users (e.g. strollers). Remarkably, even electric scooters and oncoming traffic *while the ego-vehicle is driving* are recognized as salient sound sources. A key observation from Figure 1c is that sounds originating behind corners reflect in particular patterns on nearby walls. Overall, these results show the feasibility of acoustic detection of (occluded) traffic.

B. Non-line-of-sight dataset and evaluation metrics

The quantitative experiments are designed to separately control and study various factors that could influence acoustic perception. We collected multiple recordings of the situations explained in Section III at five T-junction locations with blind corners in the inner city of Delft. The locations are categorized into two types of walled acoustical environments, namely types A and B (see Figure 5). At these locations common background noise, such as construction sites and other traffic, was present at various volumes. For safety and control, we did not record in the presence of other motorized traffic on the roads at the target junction.

The recordings can further be divided into *Static* data, made while is the ego-vehicle in front of the junction but not moving, and more challenging *Dynamic* data where the ego-vehicle reaches the junction at ∼15 km/h (see the supplementary video). Static data is easily collected, and ensures that the main source of variance is the approaching vehicle's changing position.

For the static case, the ego-vehicle was positioned such that the building corners are still visible in the camera and occlude the view onto the intersecting road (on average a distance of ∼7-10m from the intersection). Different types of passing vehicles were recorded, although in most recordings the approaching vehicle was a Škoda Fabia 1.2 TSI (2010) driven by one of the authors. For the Dynamic case, coordinated recordings with the Škoda Fabia were conducted to ensure that encounters were relevant and executed in a safe manner. Situations with `left`/`right`/`none` approaching vehicles were performed in arbitrary order to prevent undesirable correlation of background noise to some class labels. In ∼70% of the total Dynamic recordings and ∼19.5% of the total Static recordings, the ego-vehicle's noisy internal combustion engine was running to charge its battery.

TABLE I
SAMPLES PER SUBSET. IN THE ID, S/D INDICATES STATIC/DYNAMIC EGO-VEHICLE, A/B THE ENVIRONMENT TYPE (SEE FIGURE 5).

ID	`left`	`front`	`right`	`none`	Sum
SA1 / DA1	14 / 19	30 / 38	16 / 19	30 / 37	90/113
SA2 / DA2	22 / 7	41 / 15	19 / 8	49 / 13	131/ 43
SB1 / DB1	17 / 18	41 / 36	24 / 18	32 / 35	114/107
SB2 / DB2	28 / 10	55 / 22	27 / 12	43 / 22	153/ 66
SB3 / DB3	22 / 19	45 / 38	23 / 19	45 / 36	135/112
SAB / DAB	103/ 73	212/149	109/ 76	199/143	*623/441*

a) Sample extraction: For each Static recording with an approaching target vehicle, the time t_0 is manually annotated as the moment when the approaching vehicle enters direct line-of-sight. Since the quality of our t_0 estimate is bounded by the ego-vehicle's camera frame rate (10 Hz), we conservatively regard the last image *before* the incoming vehicle is visible as t_0. Thus, there is no line-of-sight at $t \leq t_0$. At $t > t_0$ the vehicle is considered visible, even though it might only be a fraction of the body. For the Dynamic data, this annotation is not feasible as the approaching car may be in direct line-of-sight, yet outside the limited field-of-view of the front-facing camera as the ego-vehicle has

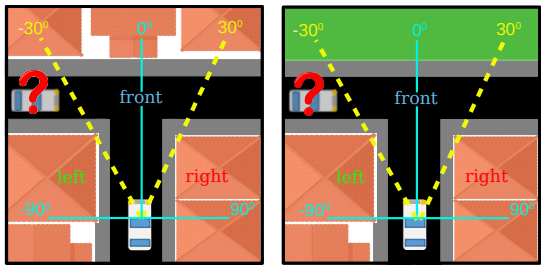
(a) Stroller at a distance (b) Electric scooter (c) Scooter overtaking (d) Car passing by (e) Oncoming car

Fig. 4. Qualitative examples of 2D Direction-of-Arrival estimation overlaid on the camera image (zoomed). (a): Stroller wheels are picked up even at a distance. (b), (c): Both conventional and more quiet electric scooters are detected. (d): The loudest sound of a passing vehicle is typically the road contact of the individual tires. (e): Even when the ego-vehicle drives at $\sim$ 30 km/h, oncoming moving vehicles are still registered as salient sound sources.

advanced onto the intersection. Thus, annotating t_0 based on the camera images is not representative for line-of-sight detection. To still compare our results across locations, we manually annotate the time τ_0, the moment when the ego-vehicle is at the same position as in the corresponding Static recordings. All Dynamic recordings are aligned to that time as it represents the moment where the ego-vehicle should make a classification decision, irrespective if an approaching vehicle is about to enter line-of-sight or still further away.

From the recordings, short $\delta t = 1$s audio samples are extracted. Let t_e, the end of the time window $[t_e - 1\text{s}, t_e]$, denote a sample's time stamp at which a prediction could be made. For Static `left` and `right` recordings, samples with the corresponding class label are extracted at $t_e = t_0$. For Dynamic recordings, `left` and `right` samples are extracted at $t_e = \tau_0 + 0.5$s. This ensures that during the 1s window the ego-vehicle is on average close to its position in the Static recordings. In both types of recordings, `front` samples are extracted 1.5s after the `left`/`right` samples, e.g. $t_e = t_0 + 1.5$s. Class `none` samples were from recordings with no approaching vehicles. Table I lists statistics of the extracted samples at each recording location.

b) Data augmentation: Table I shows that the data acquisition scheme produced imbalanced class ratios, with about half the samples for `left`, `right` compared to `front`, `none`. Our experiments therefore explore *data augmentation*. By exploiting the symmetry of the angular DoA bins, augmentation will double the `right` and `left` class samples by reversing the azimuth bin order in all $\boldsymbol{r}_l$, resulting in new features for the opposite label, i.e. as if additional data was collected at mirrored locations. Augmentation is a training strategy only, and thus not applied to test data to keep results comparable, and distinct for `left` and `right`.



(a) **Type A**: completely walled (b) **Type B**: walled exit

Fig. 5. Schematics of considered environment types. The ego-vehicle approaches the junction from the bottom. Another vehicle might approach behind the left or right blind corner. Dashed lines indicate the camera FoV.

c) Metrics: We report the overall accuracy, and the per-class Jaccard index (a.k.a. Intersection-over-Union) as a robust measure of one-vs-all performance. First, for each class c the True Positives/Negatives (TP_c/TN_c), and False Positives/Negatives (FP_c/FN_c) are computed, treating target class c as positive and the other three classes jointly as negative. Given the total number of test samples N, the overall accuracy is then $\left(\sum_{c\in\mathcal{C}} TP_c\right)/N$ and the per-class Jaccard index is $J_c = TP_c/(TP_c + FP_c + FN_c)$.

TABLE II

BASELINE COMPARISON AND HYPERPARAMETER STUDY W.R.T. OUR REFERENCE CONFIGURATION: SVM $\lambda = 1$, $\delta t = 1$, $L = 2$, DATA AUGMENTATION. RESULTS ON STATIC DATA. * DENOTES *our* PIPELINE.

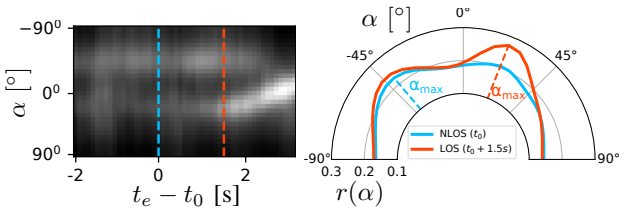
Run	Accuracy	$J_{\texttt{left}}$	$J_{\texttt{front}}$	$J_{\texttt{right}}$	$J_{\texttt{none}}$
* *(reference)*	**0.92**	0.79	0.89	**0.87**	0.83
* wo. data augment.	**0.92**	0.75	0.91	0.78	0.83
* w. $\delta t = 0.5$s	0.91	0.75	0.89	**0.87**	0.82
* w. $L = 1$	0.86	0.64	0.87	0.73	0.79
* w. $L = 3$	**0.92**	0.74	0.92	0.82	0.81
* w. $L = 4$	0.90	0.72	0.90	0.77	0.83
* w. SVM $\lambda = 0.1$	0.91	0.78	0.89	0.81	0.82
* w. SVM $\lambda = 10$	0.91	**0.81**	0.86	0.84	0.83
DoA-only [11], [9]	0.64	0.11	0.83	0.28	-
Faster R-CNN [37]	0.60	0.00	**0.99**	0.00	**0.98**

C. Training and impact of classifier and features

First, the overall system performance and hyperparameters are evaluated on all Static data from both type A and B locations (i.e. subset ID 'SAB') using 5-fold cross-validation. The folds are fixed once for all experiments, with the training samples of each class equally distributed among folds.

We fix the frequency range to $f_{min} = 50$Hz, $f_{max} = 1500$Hz, and the number of azimuth bins to $B = 30$ (Section III-B). For efficiency and robustness, a linear Support Vector Machine (SVM) is used with $l2-$regularization weighted by hyperparameter λ. Other hyperparameters to explore include the sample length $\delta t \in \{0.5\text{s}, 1\text{s}\}$, the segment count $L \in \{1, 2, 3, 4\}$, and using/not using data augmentation.

Our final choice and reference is the SVM with $\lambda = 1$, $\delta t = 1$s, $L = 2$, and data augmentation. Table II shows the results for changing these parameter choices. The overall accuracy for all these hyperparameters choices is



(a) DoA energy over time (b) DoA relative to ego-vehicle

Fig. 6. DoA energy over time for the recording shown in Figure 1c. When the approaching vehicle is not in line-of-sight (NLOS), e.g. at t_0, the main peak is a reflection on the wall ($\alpha_{max} < -30°$) opposite of that vehicle.

mostly similar, though per-class performance does differ. Our reference achieves top accuracy, while also performing well on both `left` and `right`. We keep its hyperparameters for all following experiments.

The table also shows the results of the DoA-only baseline explained in Section III-A using $\alpha_{\text{th}} = 50°$, which was found through a grid search in the range $[0°, 90°]$. As expected, the DoA-only baseline [11], [9] shows weak performance for all metrics. While the sound source is occluded, the most salient sound direction does not represent its origin, but its reflection on the opposite wall (see Figure 1). The temporal evolution of the full DoA energy for a car approaching from the `right` is shown in Figure 6. When it is still occluded at t_0, there are multiple peaks and the most salient one is a reflection on the left ($\alpha_{max} \approx -40°$). Only once the car is in line-of-sight ($t_0 + 1.5$s) the main mode clearly represents its true direction ($\alpha_{max} \approx +25°$). The left and right image in Figure 1c also show such peaks at t_0 and $t_0 + 1.5$s, respectively.

The bottom row of the table shows the visual baseline, a Faster R-CNN R50-C4 model trained on the COCO dataset [37]. To avoid false positive detections, we set the score threshold of 75% and additionally required a bounding box height of 100 pixels to ignore cars far away in the background, which were not of interest. Generally this threshold is already exceeded once the hood of the approaching car is visible. While performing well on `front` and `none`, this visual baseline shows poor overall accuracy as it is physically incapable of classifying `left` and `right`.

D. Detection time before appearance

Ultimately, the goal is to know whether our acoustic method can detect approaching vehicles earlier than the state-of-the-art visual baseline. For this purpose, their online performance is compared next.

The static recordings are divided into a fixed training (328 recordings) and test (83 recordings) split, stratified to adequately represent labels and locations. The training was conducted as in Section IV-C with `left` and `right` samples extracted at $t_e = t_0$. The visual baseline is evaluated on every camera frame (10 Hz). Our detector is evaluated on a sliding window of $1s$ across the 83 test recordings. To account for the transition period when the car may still be partly occluded, `front` predictions by both methods are accepted as correct starting at $t = t_0$. For recordings of classes `left` and `right`, these classes are accepted until $t = t_0 + 1.5$s, allowing for temporal overlap with `front`.

Figure 7 illustrates the accuracy on the test recordings for different evaluation times t_e. The overlap region is indicated by the gray area after $t_e = t_0$ and its beginning thus marks when a car enters the field of view. At $t_e = t_0$, just before entering the view of the camera, the approaching car can be detected with 0.94 accuracy by our method. This accuracy is achieved more than one second ahead of the visual baseline, showing that our acoustic detection gives the ego-vehicle additional reaction time. After 1.5s a decreasing accuracy is reported, since the leaving vehicle is not annotated and only `front` predictions are considered true positives. The acoustic detector sometimes still predicts `left`, or `right` once the car crossed over. The Faster R-CNN accuracy also decreases: after 2s the car is often completely occluded again.

Figure 8 shows the per-class probabilities as a function of extraction time t_e on the test set, separated by recording situations. The SVM class probabilities are obtained with the method in [38]. The probabilities for `left` show that on average the model initially predicts that no car is approaching. Towards t_0, the `none` class becomes less likely and the model increasingly favors the correct `left` class. A short time after t_0, the prediction flips to the `front` class, and eventually switches to `right` as the car leaves line-of-sight. Similar (mirrored) behavior is observed for vehicles approaching from the right. The probabilities of `left`/`right` rise until the approaching vehicle is almost in line-of-sight, which corresponds to the extraction time of the training samples. The `none` class is constantly predicted as likeliest when no vehicle is approaching. Overall, the prediction matches the events of the recorded situations remarkably well.

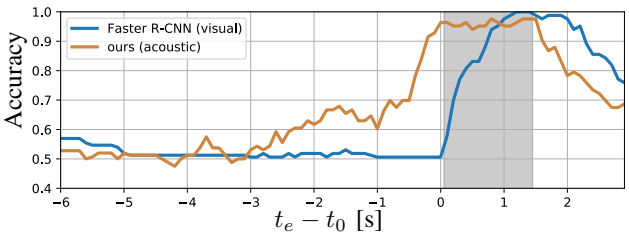


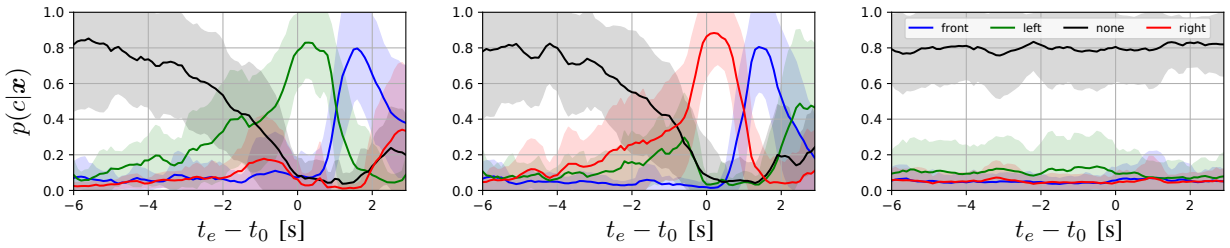
Fig. 7. Accuracy over test time t_e of our acoustic and the visual baseline on 83 Static recordings. Gray region indicates the other vehicle is half-occluded and two labels, `front` and either `left` or `right`, are considered correct.

TABLE III
CROSS-VALIDATION RESULTS PER ENVIRONMENT ON DYNAMIC DATA.

Subset	Accuracy	$J_{\texttt{left}}$	$J_{\texttt{front}}$	$J_{\texttt{right}}$	$J_{\texttt{none}}$
DAB	0.76	0.41	0.80	0.44	0.65
DA	0.84	0.66	0.85	0.64	0.72
DB	0.75	0.33	0.81	0.42	0.64

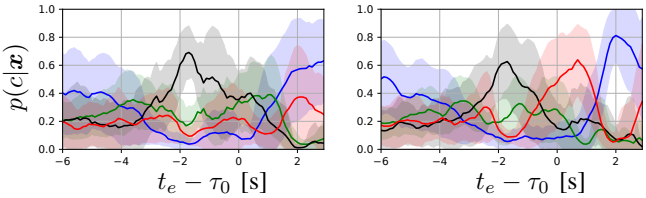
E. Impact of the moving ego-vehicle

Next, our classifier is evaluated by cross-validation per environment subset, as well as on the full Dynamic data. As



(a) `left` recordings (b) `right` recordings (c) `none` recordings

Fig. 8. Mean and std. dev. of predicted class probabilities at different times t_e on test set recordings of the Static data (blue is `front`, green is `left`, red is `right`, and black is `none`). Each figure shows recordings of a different situation. The approaching vehicle appears in view just after $t_e - t_0 = 0$.



(a) `left` recordings (b) `right` recordings

Fig. 9. Mean and std. dev. of predicted class probabilities at different times t_e on `left` and `right` test set recordings of the Dynamic data. The ego-vehicle reached the location of training data when $t_e - \tau_0 = 0.5$s.

for the Static data, 5-fold cross-validation is applied to each subset, keeping the class distribution balanced across folds.

Table III lists the corresponding metrics for each subset. On the full Dynamic data (DAB), the accuracy indicates decent performance, but the metrics for `left` and `right` classes are much worse compared to the Static results in Table II. Separating subsets DA and DB reveals that the performance is highly dependent on the environment type. In fact, even with limited training data and large data variance from a driving ego-vehicle, we obtain decent classification performance on type A environments, and we notice that low `left` and `right` performance mainly results from type B environments. We hypothesize that the more confined type A environments reflect more target sounds and are better shielded from potential noise sources.

We also analyze the temporal behavior of our method on Dynamic data. Unfortunately, a fair comparison with a visual baseline is not possible: the ego-vehicle often reaches the intersection early, and the approaching vehicle is within line-of-sight but still outside the front-facing camera's field of view (cf. τ_0 extraction in Section IV-B). Yet, the evolution of the predicted probabilities can be compared to those on the Static data in Section IV-D. Figure 9 illustrates the average predicted probabilities over 59 Dynamic test set recordings from all locations, after training on samples from the remaining 233 recordings. The classifier on average correctly predicts `right` samples (Figure 9b), between $t_e = \tau_0$ to $t_e = \tau_0 + 0.5$s. Of the `left` recordings at these times, many are falsely predicted as `none`, only few are confused with `right`. Furthermore, the changing ego-perspective of the vehicle results in alternating DoA-energy directions and thus class predictions, compared to the Static results in Figure 8. This indicates that it might help to include the ego-vehicle's relative position as an additional feature, and obtain more varied training data to cover the positional variations.

F. Generalization across acoustic environments

We here study how the performance is affected when the classifier is trained on all samples from one environment type and evaluated on all samples of the other type. In Table IV, combinations of training and test sets are listed. Compared to the results for Static and Dynamic data (see Tables II and III), the reported results in the table show a general trend: If the classifier is trained on one environment and tested on the other, it performs worse than when samples of the same location are used. In particular, the classifier trained on SB and tested on SA is not able to correctly classify samples of `left` and `right` while inverse training and testing performs much better. On the Dynamic data, such pronounced effects are not visible, but overall the accuracy decreases compared to the Static data. In summary, the reflection patterns vary from one environment to another, yet at some locations the patterns appear more distinct and robust than those at others.

TABLE IV

GENERALIZATION ACROSS LOCATIONS AND ENVIRONMENTS.

Training	Test	Accuracy	$J_{\texttt{left}}$	$J_{\texttt{front}}$	$J_{\texttt{right}}$	$J_{\texttt{none}}$
SB	SA	0.66	0.03	0.66	0.03	0.62
SA	SB	0.79	0.42	0.82	0.61	0.67
DB	DA	0.53	0.16	0.70	0.25	0.16
DA	DB	0.56	0.21	0.50	0.29	0.46

G. Microphone array configuration

Our array with 56 microphones enables evaluation of different spatial configurations with $M < 56$. For various subsets of M microphones, we randomly sample 100 out of $\binom{56}{M}$ possible microphone configurations, and cross-validate on the Static data. Interestingly, the best configuration with $M = 7$ already achieves similar accuracy as with $M = 56$. With $M = 2/3$ the accuracy is already $0.82/0.89$, but with worse performance on `left` and `right`. Large variance between samples highlights the importance of a thorough search of spatial configurations. Reducing M also leads to faster inference time, specifically 0.24/0.14/0.04s for $M = 56/28/14$ using our unoptimized implementation.

V. Conclusions

We showed that a vehicle mounted microphone array can be used to acoustically detect approaching vehicles behind blind corners from their wall reflections. In our experimental setup, our method achieved an accuracy of 0.92 on the 4-class hidden car classification task for a static ego-vehicle, and up to 0.84 in some environments while driving. An approaching vehicle was detected with the same accuracy as our visual baseline already more than one second ahead, a crucial advantage in such critical situations.

While these initial findings are encouraging, our results have several limitations. The experiments included only few locations and few different oncoming vehicles, and while our method performed well on one environment, it had difficulties on the other, and did not perform reliably in unseen test environments. To expand the applicability, we expect that more representative data is needed to capture a broad variety of environments, vehicle positions and velocities, and the presence of multiple sound sources. Rather than generalizing across environments, additional input from map data or other sensor measurements could help to discriminate acoustic environments and to classify the reflection patterns accordingly. More data also enables end-to-end learning of low-level features, potentially capturing cues our DoA-based approach currently ignores (e.g. Doppler, sound volume), and perform multi-source detection and classification in one pass [30]. Ideally a suitable self-supervised learning scheme is developed [31], though a key challenge is that actual occluded sources cannot immediately be visually detected.

References

[1] C. G. Keller, T. Dang, H. Fritz, A. Joos, C. Rabe, and D. M. Gavrila, "Active pedestrian safety by automatic braking and evasive steering," *IEEE T-ITS*, vol. 12, no. 4, pp. 1292–1304, 2011.

[2] Z. MacHardy, A. Khan, K. Obana, and S. Iwashina, "V2X access technologies: Regulation, research, and remaining challenges," *IEEE Comm. Surveys & Tutorials*, vol. 20, no. 3, pp. 1858–1877, 2018.

[3] S. Argentieri, P. Danes, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech & Language*, vol. 34, no. 1, pp. 87–112, 2015.

[4] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics & Autonomous Systems*, vol. 96, pp. 184–210, 2017.

[5] L. Wang and A. Cavallaro, "Acoustic sensing from a multi-rotor drone," *IEEE Sensors Journal*, vol. 18, no. 11, pp. 4570–4582, 2018.

[6] D. B. Lindell, G. Wetzstein, and V. Koltun, "Acoustic non-line-of-sight imaging," in *Proc. of IEEE CVPR*, 2019, pp. 6780–6789.

[7] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *IEEE/RSJ IROS*. IEEE, 2012, pp. 3288–3293.

[8] I. An, M. Son, D. Manocha, and S.-e. Yoon, "Reflection-aware sound source localization," in *ICRA*. IEEE, 2018, pp. 66–73.

[9] Y. Jang, J. Kim, and J. Kim, "The development of the vehicle sound source localization system," in *APSIPA*. IEEE, 2015, pp. 1241–1244.

[10] A. Stelling-Kończak, M. Hagenzieker, and B. V. Wee, "Traffic sounds and cycling safety: The use of electronic devices by cyclists and the quietness of hybrid and electric cars," *Transport Reviews*, vol. 35, no. 4, pp. 422–444, 2015.

[11] M. Mizumachi, A. Kaminuma, N. Ono, and S. Ando, "Robust sensing of approaching vehicles relying on acoustic cues," *Sensors*, vol. 14, no. 6, pp. 9546–9561, 2014.

[12] A. V. Padmanabhan, H. Ravichandran, *et al.*, "Acoustics based vehicle environmental information," SAE, Tech. Rep., 2014.

[13] K. Asahi, H. Banno, O. Yamamoto, A. Ogawa, and K. Yamada, "Development and evaluation of a scheme for detecting multiple approaching vehicles through acoustic sensing," in *IV Symposium*. IEEE, 2011, pp. 119–123.

[14] V. Singh, K. E. Knisely, *et al.*, "Non-line-of-sight sound source localization using matched-field processing," *J. of the Acoustical Society of America*, vol. 131, no. 1, pp. 292–302, 2012.

[15] T. Toyoda, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Traffic monitoring with ad-hoc microphone array," in *Int. Workshop on Acoustic Signal Enhancement*. IEEE, 2014, pp. 318–322.

[16] S. Ishida, J. Kajimura, M. Uchino, S. Tagashira, and A. Fukuda, "SAVeD: Acoustic vehicle detector with speed estimation capable of sequential vehicle detection," in *ITSC*. IEEE, 2018, pp. 906–912.

[17] U. Sandberg, L. Goubert, and P. Mioduszewski, "Are vehicles driven in electric mode so quiet that they need acoustic warning signals," in *Int. Congress on Acoustics*, 2010.

[18] L. M. Iversen and R. S. H. Skov, "Measurement of noise from electrical vehicles and internal combustion engine vehicles under urban driving conditions," *Euronoise*, 2015.

[19] R. Robart, E. Parizet, J.-C. Chamard, *et al.*, "eVADER: A perceptual approach to finding minimum warning sound requirements for quiet cars." in *AIA-DAGA 2013 Conference on Acoustics*, 2013.

[20] S. K. Lee, S. M. Lee, T. Shin, and M. Han, "Objective evaluation of the sound quality of the warning sound of electric vehicles with a consideration of the masking effect: Annoyance and detectability," *Int. Journal of Automotive Tech.*, vol. 18, no. 4, pp. 699–705, 2017.

[21] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *ICASSP*. IEEE, 2018, pp. 351–355.

[22] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University Providence, RI, 2000.

[23] M. Hornikx and J. Forssén, "Modelling of sound propagation to three-dimensional urban courtyards using the extended Fourier pstd method," *Applied Acoustics*, vol. 72, no. 9, pp. 665–676, 2011.

[24] W. Zhang, P. N. Samarasinghe, H. Chen, and T. D. Abhayapala, "Surround by sound: A review of spatial audio recording and reproduction," *Applied Sciences*, vol. 7, no. 5, p. 532, 2017.

[25] K. Osako, Y. Mitsufuji, *et al.*, "Supervised monaural source separation based on autoencoders," in *ICASSP*. IEEE, 2017, pp. 11–15.

[26] A. Saxena and A. Y. Ng, "Learning sound location from a single microphone," in *ICRA*. IEEE, 2009, pp. 1737–1742.

[27] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[28] A. Valada, L. Spinello, and W. Burgard, "Deep feature learning for acoustics-based terrain classification," in *Robotics Research*. Springer, 2018, pp. 21–37.

[29] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *J. of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[30] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *ICRA*. IEEE, 2018, pp. 74–79.

[31] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proc. of ICCV*, 2019.

[32] N. Scheiner, F. Kraus, F. Wei, *et al.*, "Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar," in *Proc. of IEEE CVPR*, 2020, pp. 2068–2077.

[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[34] L. Ferranti, B. Brito, E. Pool, Y. Zheng, *et al.*, "SafeVRU: A research platform for the interaction of self-driving vehicles with vulnerable road users," in *IV Symposium*. IEEE, 2019, pp. 1660–1666.

[35] F. Pedregosa, G. Varoquaux, *et al.*, "Scikit-learn: Machine learning in python," *JMLR*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[36] E. Sarradj and G. Herold, "A python framework for microphone array data processing," *Applied Acoustics*, vol. 116, pp. 50–58, 2017.

[37] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[38] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *JMLR*, vol. 5, no. Aug, pp. 975–1005, 2004.