Hearing What You Cannot See

Acoustic Vehicle Detection Around Corners

Schulz, Yannick; Mattar, Avinash Kini; Hehn, Thomas; Kooij, Julian

DOI
10.1109/LRA.2021.3062254

Publication date
2021

Document Version
Accepted author manuscript

Published in
IEEE Robotics and Automation Letters

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.**

***Index Terms*—Robot Audition; Intelligent Transportation Systems; Object Detection, Segmentation and Categorization**

I. INTRODUCTION

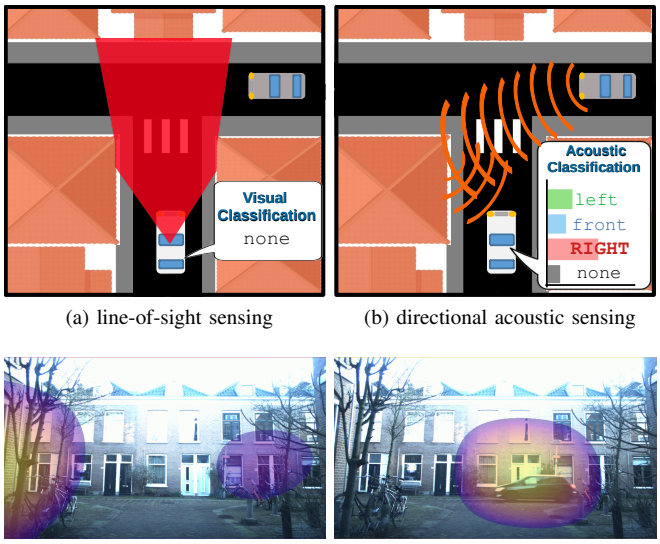
HIGHLY automated and self-driving vehicles currently rely on three complementary main sensors to identify visible objects, namely camera, lidar, and radar. However, the capabilities of these conventional sensors can be limited in urban environments when sight is obstructed by narrow streets, trees, parked 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.

Manuscript received: October, 15, 2020; Revised January, 9, 2021; Accepted February, 7, 2021.

This paper was recommended for publication by Editor Cesar Cadena upon evaluation of the Associate Editor and Reviewers' comments.

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

Digital Object Identifier (DOI): see top of this page.



(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.

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.

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 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

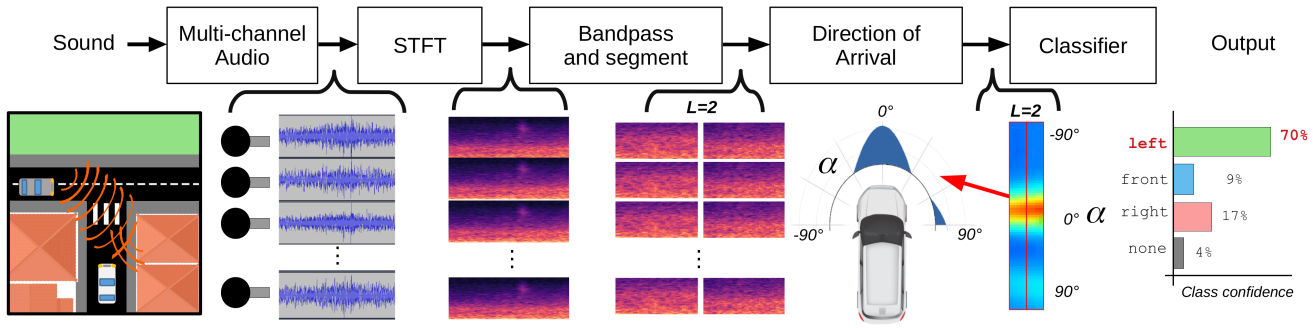


Fig. 2. Overview of our acoustic detection pipeline, see Section III-B for an explanation of the steps.

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

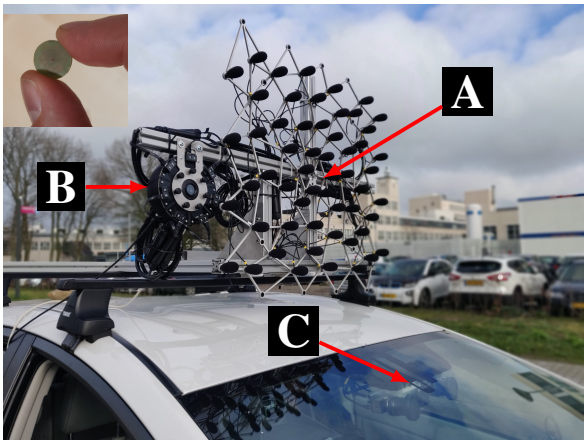


Fig. 3. Sensor setup of our test vehicle. A: Center of the 56 MEMS acoustic array. B: signal processing unit. C: front camera behind windscreen. Inset: the diameter of a single MEMS microphone is only 12mm.

(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.

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.

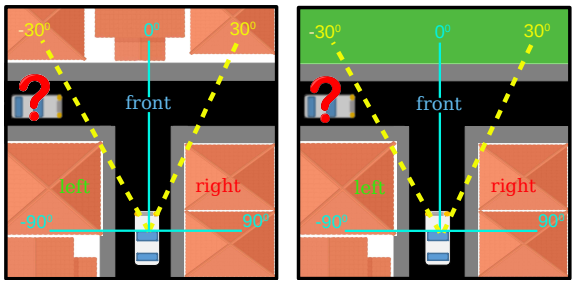
A. Line-of-sight localization – qualitative results

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 $\sim$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.



(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.

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 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 $\delta t = 1\text{s}$ 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 $t_e = \tau_0 + 0.5\text{s}$. 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. $t_e = t_0 + 1.5\text{s}$. 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`.

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.

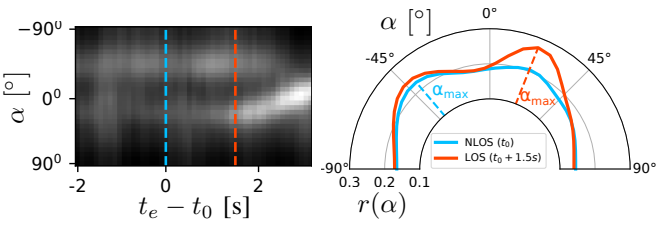
Run	Accuracy	$J_{\texttt{left}}$	$J_{\texttt{front}}$	$J_{\texttt{right}}$	$J_{\texttt{none}}$
* *(reference)*	**0.92**	0.79	0.89	**0.87**	0.83
* wo. data augment.	**0.92**	0.75	0.91	0.78	0.83
* w. $\delta t = 0.5$s	0.91	0.75	0.89	**0.87**	0.82
* w. $L = 1$	0.86	0.64	0.87	0.73	0.79
* w. $L = 3$	**0.92**	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 $f_{min} = 50\text{Hz}, f_{max} = 1500\text{Hz}$, 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 $\lambda = 1$, $\delta t = 1\text{s}$, $L = 2$, and data augmentation. Table II shows the results for changing these parameter choices. The overall accuracy for all these hyperparameters choices is 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.



(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.

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.

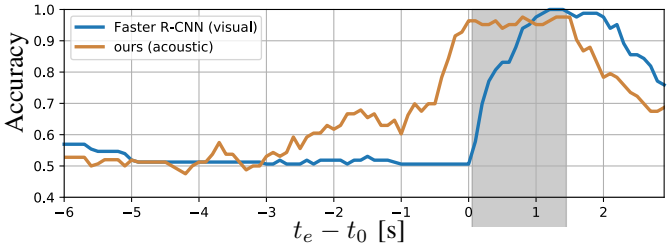


Fig. 7. Accuracy over test time t_e of our acoustic and the visual baseline on 83 Static recordings. Gray region indicates the other vehicle is half-occluded and two labels, `front` and either `left` or `right`, are considered correct.

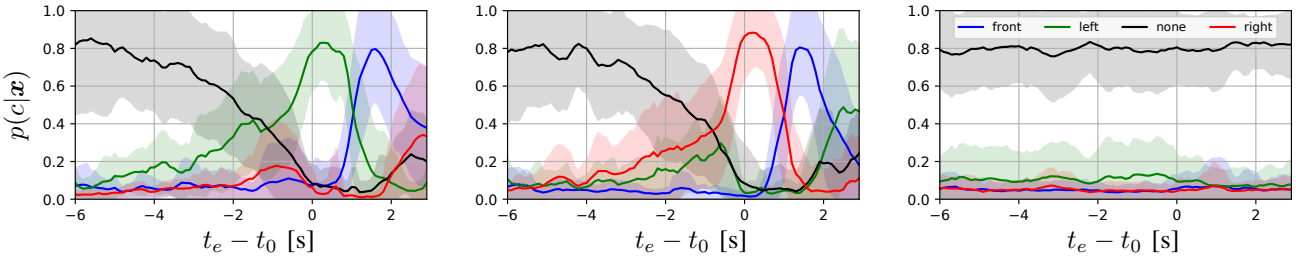
TABLE III
CROSS-VALIDATION RESULTS PER ENVIRONMENT ON DYNAMIC DATA.

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

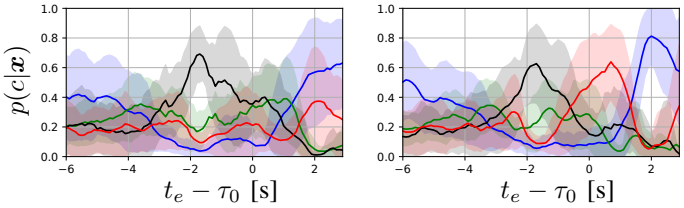
Next, our classifier is evaluated by cross-validation per environment subset, as well as on the full Dynamic data. As for the Static data, 5-fold cross-validation is applied to each subset, keeping the class distribution balanced across folds.

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`



(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.

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.