How Well Can Driverless Vehicles Hear? A Gentle Introduction to Auditory Perception for Autonomous and Smart Vehicles

Letizia Marchegiani, *Member, IEEE,* and Xenofon Fafoutis, *Senior Member, IEEE*

***Abstract*—From sirens to lane markings, the urban environment is full of sounds that are designed to navigate the attention of the driver towards events that require special care. Microphone-equipped autonomous vehicles can also use these acoustic cues for increasing safety and performance. This article explores auditory perception in the context of autonomous driving and smart vehicles in general, examining the potential of exploiting acoustic cues in driverless vehicle technology. With a journey through the literature, we discuss various applications of auditory perception in driverless vehicles, ranging from the identification and localisation of external acoustic objects to leveraging ego-noise for motion estimation and engine fault detection. In addition to solutions already proposed in the literature, we also point out directions for further investigations, focusing in particular on parallel studies in the areas of acoustics and audio signal processing that demonstrate the potential for improving the performance of driverless cars.**

***Index Terms*—Acoustic Signal Processing, Autonomous Systems, Autonomous Vehicles, Machine Learning**

I. Introduction

DRIVERLESS vehicles will be a reality soon, as the major car manufacturers envision to reach the technology necessary for full autonomy within the next decade [67, 68].

Autonomous driving (navigation) is largely based on information extracted from sensory data (perception), as illustrated in Figure 1. Indeed, reports from car manufacturers [1, 5] suggest that the vehicle's perception is implemented mainly through cameras and long- and short-range sensors, such as lasers and radars, as well as through information received from their environment [22]. Besides a few notable exceptions, such as Waymo (the successor of the Google Self-Driving Car) [63], driverless cars are rarely provided with auditory sensing. Research outputs from the academic world follow a similar pattern. Nevertheless, it is indisputable that acoustic stimuli play an important role in the understanding of certain dynamics that characterise urban environments, either because there are specific cases where no other sensing modality can, indeed, replace hearing capabilities (*e.g.* a car honking), or because having such additional information can greatly ease the interpretation of environmental cues and the generation of appropriate strategies (*e.g.* detecting the presence of an emergency vehicle approaching an intersection long before it is actually possible to see it, for instance, would increase safety).

L.Marchegiani is with the Department of Electronic Systems, University of Aalborg, Aalborg Ø, DK (e-mail: lm@es.aau.dk).

X. Fafoutis is with DTU Compute, Technical University of Denmark, 2800 Kgs. Lyngby, DK (e-mail:xefa@dtu.dk).

Although audio perception is not absent in smart vehicles, we argue that it is not used in its full potential. Generally speaking, traditional sensors, such as cameras and lasers, provide, in many contexts, sufficient information about the surrounding environment. Therefore, it is important to highlight that this article does not suggest that audio sensors should replace traditional sensing modalities. Instead, we invite the reader to consider auditory perception as a complementary sensing modality. Indeed, autonomous vehicles are safety-critical systems, and as a result, such redundancy is important for enhancing safety. Specifically, incorporating auditory sensing in a multi-modal manner (Figure 1) complements traditional sensors in two ways, namely, it reduces uncertainty and it provides a fall-back in case of subsystem failures. All sensor modalities, in fact, are affected by different limitations. Cameras are particularly sensitive to scene illumination and structure. Lasers do not cope well with harsh weather conditions, such as heavy rain, fog or snow. Similarly, audio-based systems suffer in cases of high background noise, such as in especially windy conditions, and are not able to capture events that lack an audio signature. On the other hand, audio is resilient to scene appearance, and, differently from lasers and cameras, can perceive events which are out of the field of view.

Auditory perception has been used in the literature for a variety of applications. Straightforward examples are the detection and identification of anomalous sounds, such as sirens or horns, as well as audio-based vehicle classification. Besides the opportunity of enabling or enhancing the identification of specific acoustic objects in the scene, audio signals could support driverless vehicles in performing many other tasks. For instance, previous studies have shown that the road-tyre interaction noise changes depending on road status. Similarly, road markings emit specific sounds when crossed. The ambition of realising reliable, robust and safe driverless vehicles, however, does not concern only urban on-road driving; rather it covers a variety of areas and application domains. The usage of autonomous vehicles in agriculture, for instance, would greatly facilitate the execution of several tasks, many of which often involve risks for human operators. The realisation of off-road autonomous vehicles presents additional challenges, such as localisation and control challenges [42] (*e.g.* the terrain often causes sliding and slipping, different kinds of terrain require different motion profiles, etc). Audio signals, indeed, could be very useful in these kinds of scenarios for sound-based terrain classification and to increase the robustness of the

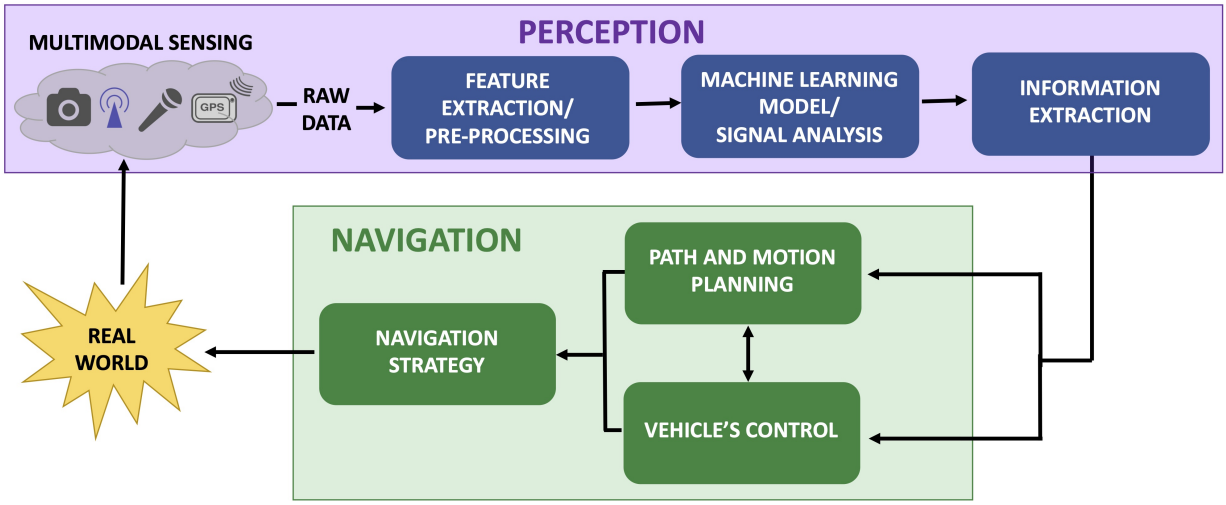


Fig. 1: Full pipeline operating in an autonomous vehicle. The environment is monitored using a variety of sensing modalities, including audio sensors. Information is extracted from the raw sensor data using machine learning and signal processing. In turn, the extracted information is used to safely navigate the vehicle. The focus of this article is on auditory perception in the context of this pipeline.

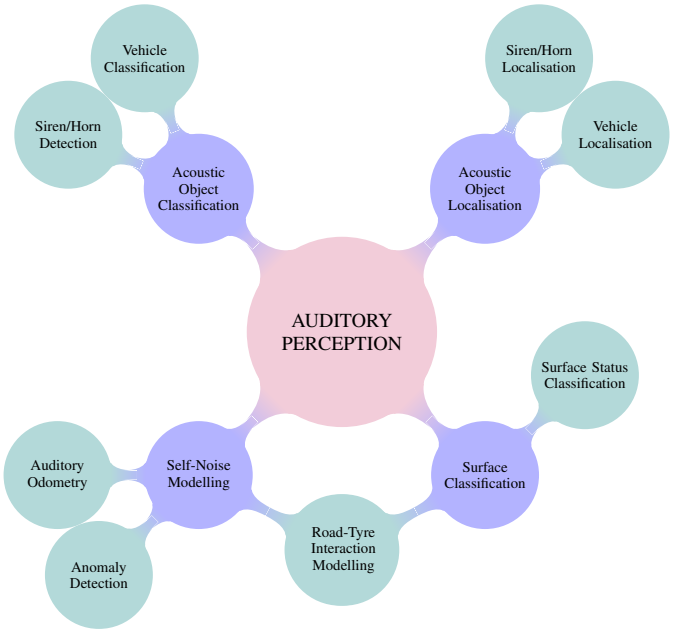


Fig. 2: Linked representation of the various areas explored in the article.

vehicle's odometry, complementing traditional sensors, which struggle in difficult lighting or weather conditions. A further application of auditory perception for smart vehicles is engine fault detection. It is apparent, indeed, that one of the first signs of defective operation of a car is caught by human drivers through anomalous ego-noise emitted by the vehicle.

This article constitutes a gentle introduction to the use of auditory perception in autonomous driving. In summary the contributions of this article are summarised as follows: (i) we begin with an illustration of the solutions proposed in the literature, focusing in particular on the employed signal processing technologies and machine learning frameworks; (ii) with a journey through the state of the art, we highlight the potential of leveraging acoustic information for autonomous driving and introduce the technological building blocks of this emerging trend; (iii) we discuss potential directions for further investigations, presenting studies in the areas of acoustic and audio signal processing that have the potential to advance the field of autonomous driving.

The remaining of this article is structured as follows: Section II analyses methods for soundscape understanding and interpretation; Section III explores road-tyre interaction noise for road status and driver behaviour monitoring; Sections IV and V approach ego-noise modelling for off-road applications and fault detection, respectively; Section VI concludes the article. Figure 2 depicts a linked representation of the research areas explored in this article, while Table I summarises the datasets used in the surveyed literature. The article provides only the necessary information on the experimental setup of the surveyed works; we invite the interested reader to find all the details in the referenced papers.

II. The Urban Soundscape

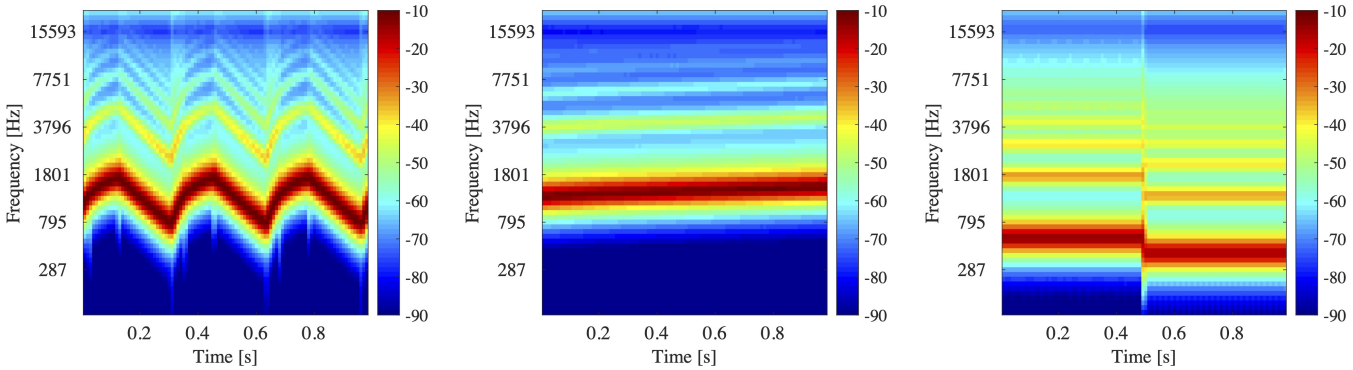
This section explores the use of auditory signals for urban environment modelling and interpretation. Specifically, we first discuss ways for detecting and classifying alerting acoustic events, such as sirens of emergency vehicles or horns; we then illustrate methods for the detection and recognition of different classes of road vehicles.

A. Alerting Sound Detection and Recognition

Sirens of emergency vehicles do not sound the same all around the world. Different countries rely on different kinds of sirens; some adopt a series of them, and then switch from one to another depending on the situation. The most common are named *yelp*, *wail*, and *hi-low* [23, 66]. The three sirens

Task	Source	Nature	Size	# Classes	Modality	Method	Evaluation Metric
Alerting Sound Detection and Localisation	[53]	Real and Simulated	N/A	2	Audio	MDF	Detection Accuracy
	[24]	Simulated	N/A	2	Audio	APNC	Absolute Error (Localisation)
	[62]	Real	1K	2	Audio	PBMs	Detection Accuracy
	[55]	On-line	1K	5	Audio	GMMs	Classification Accuracy
	[52]	Real	330K	4	Audio	k-NN	Classification Accuracy
	[51]	Real, Simulated & On-line	30K	3	Audio	DCNNs	Classification Accuracy & Absolute Localisation Error
Vehicle Classification	[13]	Real	200	3	Audio & Video	NBc	F-measure
	[31]	Real	1K	3	Audio & Video	NBc	F-measure
	[35, 36]	Real	178	5	Audio	DTW	F-measure & Speed Error
Road-Tyre Interaction	[39]	Real	2K	3	Audio	NN	Classification Accuracy
	[9]	Real	11K	2	Audio	SVMs	Classification Accuracy
	[6]	Real	800K	2	Audio	LSTM	Average Recall
Terrain Classification	[47]	Real	21K	6	Audio	SVMs	Classification Accuracy
	[64, 65]	Real	15h	9	Audio	DCNNs & LSTM	Classification Accuracy
Motion Estimation	[58]	Real	5K	11	Audio	MLP	Classification Accuracy
	[50]	Real	100K	Regression	Audio	DNNs	Absolute Error (Velocity Estimation)
Fault Detection	[49]	Real	N/A	2	Audio	Similarity Metrics	-
	[70]	Real	N/A	9	Audio	WT-NN	Classification Accuracy
	[18]	Real	150K	4	Audio	NN	Classification Accuracy
	[8]	Real	N/A	8	Piezoelectric sensor	NN	Classification Accuracy

TABLE I: Summary of datasets and employed methods. None of the datasets is publicly available.



(a) (b) (c)

Fig. 3: Example gammatonegrams of sirens: yelp (a), wail (b), hi-low (c). The colormaps represent the energy of the signal (dBFS).

are characterised by different time-frequency patterns: the yelp siren follows a "sinusoidal" shape, while the hi-low alternates between two main frequencies (one lower and one higher, from which the name), and the wail is composed of a fundamental frequency which changes linearly over time. A time-frequency representation of each of them is given in Figure 3.

One of the first approaches to automatic siren detection has been presented in [53]. The authors focus on hi-low sirens alternating between 392 Hz and 660 Hz, pointing out that, even though the task might appear simple (*i.e.* could be carried out relying on pure filtering and spectral analysis, such as a Fast Fourier Transform, FFT), the reality is more complex. Electronic siren generators, indeed, produce a square wave from a saturating push-pull type output stage, which introduces additional harmonics that result to be comparable to the ones of the intended signal. To lessen the effect of those unwanted frequencies, the authors propose the use of a pitch detection algorithm. The algorithm aims to isolate the periodic components of the sirens from the rest (*i.e.* unwanted harmonics present in the original signal and potential noise in the environment). Firstly, the pitch is estimated through peak searching, then the signal is parsed through a Module Difference Function (MDF) to discriminate between *pitched* and *unpitched* portions of the sound. The pitch detection algorithm outputs a signal representing the time evolution of the pitch. This signal is eventually classified as containing/not containing a siren, depending on the number of pitches that are inside the bands centred around the two main frequencies. It is also shown that the performance of the system can improve when a high-pass filter is applied to the original audio sample to remove some of the traffic noise (wind and car engine, for instance, are characterised by low-frequency components), easing the work of the pitch estimation algorithm.

Fazenda et al. in [24] explore potential following stages of a siren detection pipeline, targeting signal extraction and sound source localisation. The goal of this work is to equip common cars with a system that alerts the driver of the proximity of incoming emergency vehicles. Specifically, they relay the emergency signal to the passenger's cabin, recreating the direction of arrival of the sound through a multi-channel

loudspeaker system. The authors do not specify which kind of siren they aim to localise, and do not employ data collected in traffic scenarios, but generate ad-hoc experiments where the siren is overlapped with some background noise. The relative SNR level of the resulting mixture is not provided. These kinds of experiments, although not carried out in a fully realistic setting, allow for a more accurate estimation of the localisation performance of the system, which would be hard to obtain in *the wild*. The signal extraction algorithm employs an *adaptive predictor noise canceller* (APNC) scheme based on Least Mean Squared (LMS) optimisation, as proposed in [12]. The framework makes use of a microphone array consisting of four microphones disposed at 90° from each other. Signal extraction is applied to all four channels of the incoming audio signal, and two different localisation methods are applied and evaluated. Both methods are utilised to obtain horizontal localisation (*i.e.* direction of arrival of the sound on the horizon plane). Better results are obtained when employing a classic time delay estimation approach, based on generalised cross-correlation techniques [38].

One of the first attempts to siren detection using machine learning has been presented in [62]. The authors focused on German sirens, as defined in [2], which are hi-low sirens with characteristics similar to the ones analysed by [53]. The authors, however, move away from more traditional frameworks employed in audio classification tasks, such as Hidden Markov Models (HMMs) and Mel-Frequency Cepstrum Coefficients (MFCCs), as being *"rigid in the spectral dimension"*. In contrast, they propose a modification of Part-Based Models (PBMs), originally introduced for computer vision applications [25]. PBMs are flexible because they model the appearance and the configuration of the different sections of an image with Gaussian distributions. In [62], PBMs are applied to the Mel-spectrum of the signal. Experimental evaluation is carried out at different SNR levels (from clean to -20 dB), and the performance is compared to the one of a HMM-MFCC framework. Results, indeed, prove that the higher degree of modelling flexibility offered by PBMs helps the detection, compared to HMMs, especially when noise is present. The performance of the system, however, drops quite abruptly when the SNR is less than -5 dB.

A wider perspective on acoustic traffic events detection has been later offered by [51, 52, 55]. In [55], the authors analyse five classes of audio events: several types of ambulance siren, railroad crossing bell, tyre screech, car honk, and glass break. Gaussian Mixture Models (GMMs), operating on MFFCs and their respective *delta* and *delta-delta* coefficients, are used to discriminate among the acoustic events. Specifically, the behaviour of a GMM-based Universal Background Model (GMM-UBM) framework, firstly introduced for speaker verification purposes [59], is combined with the one of a GMM-supervector system. The system produces a feature representation of the audio signals in the form of a super-vector, obtained by concatenating the GMM mean vectors extracted from different audio segments, and adapted to the UBM, following a Maximum A Posteriori (MAP) approach [16]. A Probabilistic Principal Component Analysis (PPCA) model, as well as a linear discriminant analysis (LDA) projection, are, then, applied to the super-vectors for dimensionality reduction. The final classification is based on the cosine distance. The framework yields a notable classification accuracy, but nothing can be said about the level of SNR at which the modelling and evaluation take place, and, consequently, on how well the system will work in different noise conditions.

The work of [52] provides a different view, directly approaching noise removal prior to classification to improve accuracy. The authors propose a two-step method. Firstly, anomaly detection, based on One-Class Gaussian Processes [37], is applied to spot the presence of any potential alerting acoustic events. In case an acoustic event is detected, the classification of the event is carried out using a k-NN (k-Nearest Neighbour) framework. Classification is performed to discriminate among sirens (several types are considered), car horns and pedestrian traffic lights (*i.e.* accessible traffic signals). The k-NN framework is also augmented by an original method for the detection of samples, at testing time, which do not fall into any of the considered classes (*i.e.* sires, horn, pedestrian lights) to increase the general robustness and accuracy of the system. Before classification is applied, the authors introduce the concept of *Empirical Binary Masks (EBMs)*, which, similarly to Ideal Binary Masks [69], aims to remove unwanted masking signals from the noisy mixture. The EBMs are generated by applying k-means segmentation to the Gammatonegram [48] of the noisy signal. The impact of the noise removal step on the classification is demonstrated by comparing the performance of the k-NN framework when operating on the EBMs, the original noisy Gammatonegrams, and MFFCs. An example of segmentation is shown in Figure 4.

Despite the notable advantage in the classification provided by the EBM-based noise removal, and the great convenience given by the fact that segmentation is obtained in a fully unsupervised manner, the method makes some assumptions which do not always hold when the SNR gets too low. In fact, even though the clustering takes place without the need of any labels, the creation of a binary mask is based on understanding which clusters correspond to noise and which one corresponds to the target signal. In [52], the authors associate the target signal with the most powerful (greatest energy content) cluster, which is a fair assumption in many situations, considering that traffic emergency/alerting signals are designed to overcome the environmental noise and be easily heard by drivers [20]. Yet, when the SNR becomes particularly low, this assumption is no longer valid. The same can be said about the possibility of assigning clusters based on the spectral characteristics of the sounds we intend to detect, as also shown in [55]. To address these shortcomings, the same authors propose a different approach in [51]. Here, the segmentation relies on the use of deep learning, and, more specifically, of a U-net [60]. Rather than following a two-step approach, in this instance, the framework consists of a multi-task deep learning architecture which simultaneously segments the gammatonegram and classifies the audio signal. The cross-correlated segmented gammatonegrams are then fed to a Deep Convolutional Neural Network (DCNN) to estimate the direction of arrival of the sound. The authors obtained high performance, both in the classification and in the locali-

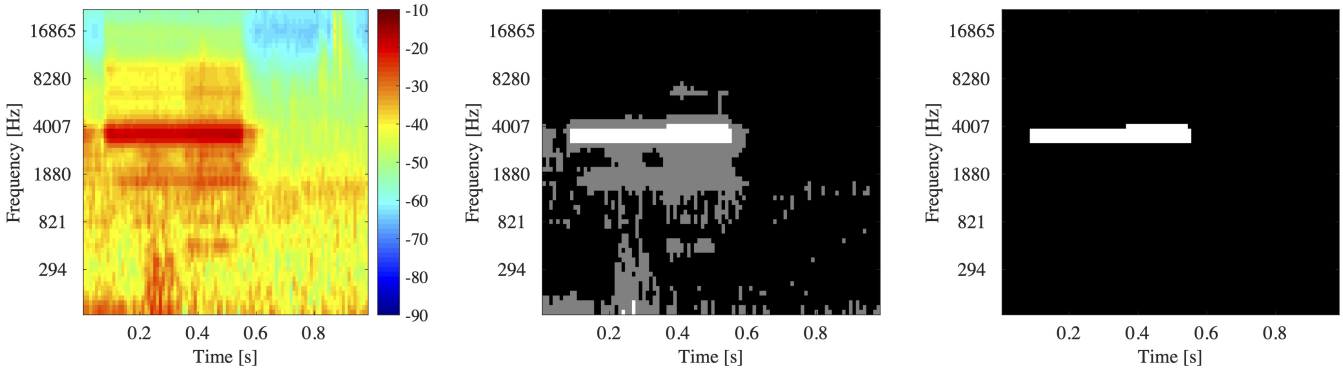


Fig. 4: Example of noise removal as proposed in [52]: The gammatonegram of a car horn in presence of traffic noise (a) is segmented into three clusters (which, intuitively, correspond to target signal, noise, and silence) using k-means (b), and finally converted into an EBM (c). The colormap on the gammatonegram represents the energy of the signal (dBFS). In the segmented gammatonegram (b) each colour indicates a different cluster.

sation tasks, even when operating in severely noisy conditions (-40 dB $\leq$ SNR ≤ 10 dB). Operating in such harsh acoustic scenarios is important for safety, as it allows the autonomous car to operate effectively in conditions whereby a human would not be able to discriminate among different acoustic events and localise the respective sound source. This is of crucial importance, especially when the vehicle is operating on a semi-autonomy regime. Several studies, indeed, have been carried out to investigate the possibility of estimating (and by doing that, anticipating) when the vehicle might fail and ask the safety driver to intervene (see [33] among others). The sooner, indeed, we are able to predict when complicated scenarios are approaching, the more time the safety drivers have to get prepared and act accordingly. An approaching emergency vehicle is a good example of those complicated scenarios, as most of the common driving rules do not apply (*e.g.* passing safely the crossroads when the traffic light is green), introducing unexpected behaviours from the drivers (*e.g.* pull over to leave the lane to the incoming emergency vehicle). Thus, the capability of a perception system to act robustly and accurately in scenarios characterised by a large presence of noise becomes crucial to guarantee safety. Nevertheless, to the authors' knowledge, no other solutions, besides [51], have addressed this issue.

B. Vehicle Classification

Human driving takes great advantage of acoustic awareness of the environment. We can perceive approaching vehicles that we cannot directly see by the sound they make while getting closer. The sound also gives us information on the direction of arrival of those vehicles and, in most of the cases, we can also discriminate among several kinds of vehicle, as they are characterised by a quite distinctive auditory signature [54]. Autonomous vehicles can also take advantage of such sounds to account for the limited field of view of visual sensors, and to complement laser-based systems [17] which are more sensitive to heavy rain, fog and snow.

The automation of vehicle classification has been explored in the literature, mainly for traffic monitoring purposes, where specific road infrastructure is in place to gather data on particular sections of streets. In [13] and [31], the authors present a multi-modal approach to discriminate between cars and trucks. Specifically, in [31] they propose audio-visual co-training of Naive Bayes classifiers (NBc), which are firstly trained on labelled samples and then labels for new samples are iteratively generated based on the confidence level of both the audio and visual models in a cooperative manner. The system is evaluated on data collected on a bridge-over where a camera and microphones record passing vehicles and provides high performance. The work presented in [13] has a similar spirit, but in this case, the audio classifier acts as an autonomous supervisor, which supports the visual on-line classifier in its continuous self-learning scheme. The process is based on boosting techniques, both for feature selection [32] and self-learning [41]. The reason behind this choice is that the sound-based model does not need a large amount of labelled training samples to start operating accurately. The framework is evaluated on real-world datasets of multi-lane freeway traffic, demonstrating great accuracy and robustness to several degrees of miss-classifications provided by the audio classifier.

Audio-only vehicle detection is provided in [36]. This work employs a stereo audio signal collected by two microphones on a road to generate a sound map from the tyre noise of passing vehicles. More specifically, the sound map is produced based on the time difference in the vehicle's sound on the two microphones. The time difference between the two sensors is expressed via generalised cross-correlation. A sound map example is illustrated in Figure 5. We can observe that each vehicle passing draws an 'S' shape on the sound map, whose orientation depends on the direction of motion of the vehicle. The authors build upon their previous work that was relying on the use of DTW (Dynamic Time Warping) [35], which, however, did not allow for sequential or simultaneous detections, as it was not able to handle overlaps in the sound

map (*i.e.* two sequential vehicles might share some points in the sound map making data association unfeasible). In [36], instead, points that correspond to an identified vehicle are removed from the sound map, minimising the interference with potential new incoming vehicles. The association between sound maps and vehicles is obtained through the application of the RANSAC (random sample consensus) algorithm [26]. The framework greatly outperforms the previous system providing high detection accuracy, also when dealing with sequential vehicles.

III. Road-Tyre Interaction

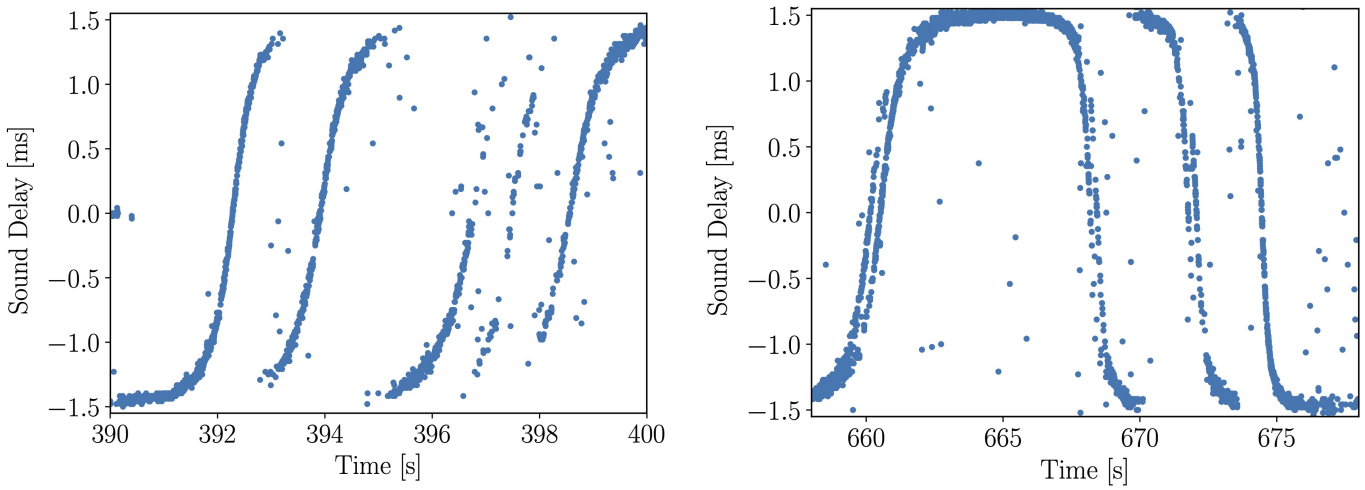
One of the first studies of road-tyre interaction noise has been reported in 1975 [40]. This work discusses, together with other parameters (*e.g.* load of the vehicle, speed), the impact of different kinds of road textures as well as surface wetness on the produced noise, expressed as the maximum A-weighted sound level[3]. Some years later, [61] carries out a more detailed analysis of the factors which might help the characterisation of road surfaces with respect to road-tyre noise interaction. Specifically, the author considers the texture profile spectrum, the sound absorption coefficient of sound propagation, and the mechanical stiffness or impedance. More recent works in this direction have been described in [21] and [30]. The former discovers the presence of a different phenomenon responsible for a decrease of the noise level in the low and medium frequencies, which operates in conjunction with an increase of the noise level in the medium and high frequencies. This phenomenon does not seem to be correlated with the speed of the vehicles, but rather with the tyre and surface characteristics. The latter reinforces the finding that the noise level is affected by the type of surface, degree of wetness, type and velocity of the vehicle. Gardziejczyk observes that a good portion of the increase in the vehicle noise can be attributed to the presence of water on the road, but that only a significant amount of water can actually set off the noise increase. The author also reports that not much difference can be observed in this context, between the noise generated by passenger cars and heavy trucks. A similar study was later performed by Freitas et al. in [27], who analysed the noise generated by a set of light and heavy vehicles on porous asphalt and dense asphalt surfaces of a motorway. Most of these studies rely on the use of the Controlled Pass-By (CPB) and the Statistical Pass-By (SPB) methods for noise measurements, as specified in the ISO 11819-1 [4]. SPB has the advantage of being applicable to normal vehicles passing by, and it does not require the employment of expensive equipment. CPB, instead, needs specified test vehicles with specified sets of market tyres. Yet, both techniques require road sensing infrastructures, and are meant to measure the amount of noise around the monitored roads.

Gail et al. in [28, 29] carried out an extensive study on the influence of surface textures of road markings on tyre-road interaction noise. They analysed seven different agglomerate road markings: irregular scattered dots, irregular dense structure, irregular lengthwise structure, regular broad drops, regular dense dots, regular narrow drops, and irregular perforate plate structure; stone mastic asphalt was utilised as a reference. Their results clearly demonstrate that, in most of the cases, road markings generate a substantial increase of the sound pressure level in the lower part of the frequency spectrum (*i.e.* $800-1000$ Hz). The authors argue that, despite the concern that such increased level of noise might cause annoyance to the residents living close to those streets, road markings play a crucial role in all those situations where out-of-lane behaviour can be considered particularly dangerous, such as tunnels, bridges etc. Human drivers rely on this noise increase, in certain cases even accompanied by vibrations, to augment their environmental awareness. It is apparent, then, that autonomous vehicles could also benefit from detecting and interpreting road markings.

One of the first attempts to the automatic classification of road status has been reported in [39]. Specifically, the authors employ a multi-modal framework which relies on the use of audio-visual information and neural networks (NN) to discriminate between wet, dry and snow-compacted surfaces. Also in this instance, the microphones were positioned on the roads' sides. On-board automatic road wetness detection has been firstly proposed in [9] and later addressed, in a deep learning perspective, in [6]. Both approaches aim to develop a warning system to improve driving safety. Nevertheless, [6] also mentions the possibility of using the output of the predictions by the machine learning frameworks in driverless vehicles to allow them to adjust the driving style and the speed profile accordingly. In [9], the microphones are placed at the front and rear of the wheels farthest from the motor engine. The authors claim that this configuration is able to minimise the impact of engine noise as well as of other forms of noise. The data collected is then parsed and fed to a Support Vector Machine (SVM) classifier for the road status detection. While most of the previously mentioned studies rely on A-weighted sound level for noise measurement, the authors here give up the A frequency weighting to better characterise the road-tyre noise level at low frequencies. Indeed, while A-weighting is effective for analysing the impact of noise on human ears, removing certain spectral components might end up being detrimental when operating with machine learning techniques. In [6], a shotgun microphone was located behind the rear tyre, and data was gathered with the car travelling at different speeds, in different traffic conditions, and pavement roughness, expanding the set of scenarios considered by [9]. The authors employ Recurrent Neural Networks (RNNs) in the form of Long-Short Term Memory (LSTM) and bi-directional LSTM (BLSTM) [34], and manage to obtain impressive accuracy, overcoming the results obtained by [9].

IV. Off-Road Applications

While the main concerns in urban autonomous driving are concentrated in the detection and safe interaction with all the other objects in the scene, such as pedestrians and other vehicles, as well as the correct interpretation and application of traffic laws, off-road scenarios present a full set of different challenges, where robust and accurate perception capabilities could be of extreme help to guarantee the correct execution of



(a) (b)

Fig. 5: Examples of sound maps, as proposed by [36]. Each vehicle yields an 'S' shape on the map. The orientation of the 'S' shape depends on the direction of motion of the vehicle: in (a) all vehicles are going in the same direction (left to right), in (b) one vehicle is going from right to left, three are going from left to right. Figures are courtesy of the authors of [36]. Figure (b) has been reproduced from a figure in [36].

several tasks. Many of the tasks, indeed, preclude any sort of human intervention, making the concept of semi-autonomy not feasible. Furthermore, dealing with a variety of terrains with diverse characteristics sheds a different light on localisation and navigation issues. Off-road vehicles might need to travel on arduous surfaces and forced to continuously adjust control mechanisms and motion profiles, for instance, to better cope with different kinds and levels of hazards [42]. Lastly, specific environments might particularly reduce the effectiveness of more traditional sensing modalities in a variety of contexts. Relying on acoustic sensing to automatically discriminate between different surfaces would be extremely helpful for the development of more informative and safe motion planning and control strategies, for example. This is especially true in all those instances, whereby sensors typically used for the same purpose might struggle to operate robustly. It is apparent, indeed, that vision-based systems do not cope well with scarce illumination, the same way lasers do not perform correctly in harsh weather conditions (*e.g.* heavy rain, fog) or in degenerated scenes, characterised by the prevalence of planar areas.

A. Terrain Classification

The idea of using proprioceptive sensors for terrain classification is not new. In 2005, vibration-based terrain analysis has been presented in [14]. This work uses an accelerometer to measure the vibration induced on a rover while approaching different terrains. Building on these results, the same authors later provide in [15] a semi-supervised approach, where proprioceptive sensors are used to train a vision-based classifier. On the side of a vibration-based framework, they also investigate the possibility of generating labels for the travelled terrains using measurements of wheel torque and sinkage to estimate the minimum traction available at the wheel–terrain interface.

One of the first works that introduces the idea of relying solely on wheel-surface interaction noise to classify the terrain the vehicle is traversing has been offered by Odedra et al. [57]. However, the first system able to perform automatic sound-based terrain classification has to be attributed to Libby and Stentz [47]. In this study, the authors build SVM classifiers to discriminate between several kinds of wheel-surface interactions. Specifically, they focus on both benign interactions, such as driving over grass, pavement and gravel roads, and hazardous terrain interactions, like splashing in the water, hitting hard objects and wheel losing traction where the terrain gets slippery. The system is tested considering a wide range of feature representations, whose impact is also analysed, obtaining a good level of recognition accuracy for most of those classes.

A similar investigation has been extended in a deep learning context in [64, 65]. In those, the authors analyse both indoor and outdoor surfaces, for a total of nine different classes of terrain: asphalt, mowed grass, medium-high grass, paving, cobblestone, off-road, wood, linoleum, and carpet. For the experimental phase, a total of 15 hours of wheel-terrain interaction noise, organised in two datasets: one dataset is gathered through a shotgun microphone mounted in the proximity of the wheels of a Pioneer P3-DX platform, and the other one through a mobile phone microphone. In [65], the classification relies on a DCNN, which yields high recognition accuracy. The DCNN is trained directly on the spectrograms of the recorded acoustic samples, and the evaluation shows that such architecture clearly outperforms methods which, instead, employ hand-crafted features. In [64], also the temporal evolution of audio signals is considered, through the introduction of an LSTM framework. All those results are particularly important, as the classes of terrain targeted and correctly recognised are characterised by a really similar visual appearance, which

would make the job of a visual-based classifier operating in the same conditions very hard, and probably lead to low performance. The capability of distinguishing among different surfaces also offers implicitly the opportunity to implement navigation systems able to detect out-of-path behaviours.

B. Motion Estimation

As mentioned earlier, harsh off-road environments might limit the performance of more classic sensors (*e.g.* cameras, lasers) when executing a variety of tasks; vehicle self-localisation represents one of the most notable examples. Self-localisation heavily relies on odometry systems, both proprioceptive (*i.e.* wheel encoders) and exteroceptive (*e.g.* cameras, lasers). A motion estimation framework able to operate despite changes in illumination, scarcity of texture, grim weather conditions (*e.g.* tick fog) or slippery terrains, would indubitably contribute to the accuracy of the whole localisation, and consequent navigation, processes. Audio-based motion estimation fits those requirements very well. An early attempt in this direction has been presented in [58], where the sound emitted by the motors (*i.e.* ego-noise) of a wheeled mobile robot is used to discriminate among the different speed profiles the platform is actually following. By relying on a Multi-Layer Perceptron (MLP) framework, the authors are able to recognise eleven different profiles, and detect modifications on the environmental conditions, such as changes in the inclination of the surface the platform is moving on. Sound-based metric motion estimation has been proposed in [50], yielding to the realisation of an *Auditory Odometry (AO)* systems. In this work, the authors use a Deep Neural Network (DNN) in a multi-task learning scheme to simultaneously regress the linear and the angular velocity at which the robot is travelling, by relying solely on the vehicle's ego-noise. They obtain great prediction accuracy, and the experimental evaluation also demonstrates significant resilience to environmental noise. It is apparent that an auditory odometry system could not alone represent a solution to robot navigation, as the robot would, basically, move "blindly". Nevertheless, multi-modal paradigms which would employ AO as one of the components for motion estimations would be able to operate more robustly, by overcoming some of the constraints of other odometry systems. A trivial example, which could very well be applied also in urban scenarios, is the way sound could assist when dealing with distractions, which often cause dramatic inaccuracies in other frameworks (*e.g.* visual odometry systems). Even a basic consensus mechanism, indeed, could, in this scenario, help the vehicle realise whether the drastic change in appearance between consecutive frames is indeed to be attributed to the motion of the vehicle or to a large modification of the environment (*i.e.* the distraction).

V. Fault Detection

The work on *Auditory Odometry* presented in [50] demonstrates that engine noise carries information which can be used to accurately estimate the motion of the vehicle. This suggests that motor noise is, indeed, quite distinctive, and could be used, in more general contexts, to analyse the behaviour of the engine itself. In 1997, a technical report from SAE International [46] proposes the idea of an audio-based engine failure diagnosis system, exploring the impact of different acoustic features for the detection of motor anomalies. The report vouches for the use of Wavelet transforms, Fast Fourier Transform (FFT) and cepstrum analysis for the identification of tappet clicks fault, engine misfiring, and abnormal combustion sound, respectively. In this spirit, an autonomous vehicle could incorporate engine fault detection to enhance the safety of autonomous driving. For example, an autonomous vehicle could monitor the noise of its engine and, if a fault is detected, drive to the emergency lane and gracefully stop. Moreover, in semi-autonomous vehicles, engine fault detection system could be used to enable safety driver intervention.

Since [46], several other studies have been carried out in the last twenty years towards the realisation of accurate sound-based fault detection frameworks. In 2009, Madain et al. [49] built a database of sounds associated to a variety of abnormal engine behaviours, and used a series of similarity metrics (*i.e.* the correlation coefficient, the normalised root mean square error, and the formant frequencies) to establish whether noise, newly generated by the motors, has to be considered as a fault or not. One year later, [70] presents a machine learning approach to the same problem, employing Wavelet transforms and neural networks (WT-NN) to discriminate among eight common engine faults. A comparative analysis of the performance of multiple neural network architectures and SVMs in the classification of air filter, spark plug, and insufficient lubricants faults is provided in [18]. By relying solely on one microphone, all the frameworks are able to operate robustly on separate faults; yet, accuracy decreases when faults appear simultaneously. An experimental evaluation of the acoustic characteristics of both normal and anomalous motor noise, at different engine speeds, is offered in [7]. More recently, Ahmed et al. [8] investigate and compare the performance of several techniques for the training of a neural network-based fault detection framework. In particular, this work focuses on backpropagation (BP), the Levenberg-Marquardt (LM) and the quasiNewton (QN) methods, the extended Kalman filter (EKF), and the smooth variable structure filter (SVSF), which provides the highest classification accuracy. A somehow complementary approach is presented in [19]. The authors target diesel engines and propose a mechanism for the extraction of fault components from abnormal sound, which can then be used for fault classification. This mechanism, which they name *Dislocation Superimposed Method (DSM)*, is based on a combination of the improved random decrement technique [10], and correlation analysis.

VI. Conclusion

We began this article with a question: *"How well can driverless vehicles hear?"*; it appears, indeed, that auditory perception is often being overlooked by the automotive industry and the autonomous driving research community. Yet, evidence suggests that the soundscape of an autonomous car is rich with information that can complement traditional sensing modalities, increasing the accuracy and safety of

Task	Benefits	Traditional Sensors	Current Limitations	Audio's Contribution
Emergency Vehicle Detection and Localisation	Emergency Navigation (e.g. stop at green light if ambulance is approaching) Alerting Safety Driver	Cameras	Limited Field of View Sensitivity to Scene Illumination	360° Field of View Resilience to Scene Illumination
		Lasers	Limited Field of View Sensitivity to Weather Conditions Ambiguous Vehicle Signature	360° Field of View Resilience to Scene Appearance Distinctive Acoustic Signature
		Radar	Ambiguous Vehicle Signature	Distinctive Acoustic Signature
Horn Detection and Localisation	Safe Vehicle Interaction Alerting Safety Driver	-	Not Perceivable	Distinctive Acoustic Signature
Vehicle Detection, Classification and Localisation	Safe Vehicle Interaction	Cameras	Limited Field of View Sensitivity to Scene Illumination	360° Field of View Resilience to Scene Illumination
		Lasers	Limited Field of View Sensitivity to Weather Conditions Ambiguous Vehicle Signature	360° Field of View Resilience to Scene Appearance Distinctive Acoustic Signature
		Radars	Ambiguous Vehicle Signature	Distinctive Acoustic Signature
Road-Tyre Interaction/ Terrain Classification	Adaptive Speed and Motion Profile	Cameras	Ambiguous Visual Signature Sensitivity to Scene Illumination	Distinctive Acoustic Signature Resilience to Scene Illumination
		Lasers	Sensitivity to Weather Conditions Ambiguous Terrain Signature	Resilience to Scene Appearance Distinctive Acoustic Signature
Motion Estimation	Self-Localisation & Navigation	Cameras	Sensitivity to Scene Structure Sensitivity to Scene Illumination Sensitivity to Distractions	Resilience to Scene Structure Resilience to Scene Illumination Resilience to Distractions
		Lasers	Sensitivity to Scene Structure Sensitivity to Weather Conditions Sensitivity to Distractions	Resilience to Scene Structure Resilience to Scene Appearance Resilience to Distractions
		Radars	Sensitivity to Scene Structure Sensitivity to Distractions	Resilience to Scene Structure Resilience to Distractions
		GPS	GPS-denied Environments	Independent of External Infrastructure
Fault Detection	Safety Navigation Procedure Alerting Safety Driver	-	Not Perceivable by Exteroceptive Sensors	Distinctive Acoustic Signature

TABLE II: Auditory perception complements traditional sensing modalities in various tasks

autonomous driving. Indeed, multiple alerting events in urban environments, such as sirens and horns, are acoustic by nature. Moreover, the engine and the interaction of the tyres on the road generate noise that can be leveraged for odometry, out-of-lane detection, and fault detection, among others. Table II summarises how auditory perception complements traditional sensors with respect to these tasks.

A. Future Directions

In this article, we explored a number of research areas whereby auditory perception has been applied to enhance autonomous and semi-autonomous vehicles. Let us now explore directions for future research.

The detection and classification of acoustic objects are fairly mature areas; yet, several of the proposed frameworks could be extended to more fine classifications. For example, vehicle classification is currently quite coarse (*i.e.* classification among cars, trucks and heavy trucks); such frameworks could be extended to identify more types of vehicles (*e.g.* motorcycles, buses) but also engine types (*e.g.* combustion vs electric engines). Furthermore, to the extent of our knowledge, there is no literature for the detection of pedestrians that is based on audio (*i.e.* speaker detection and localisation in urban scenarios). In terms of urban acoustic object localisation, on the other hand, we believe that there is room for more research: the literature is fairly limited and it is based primarily on simulated sound scenes that do not always capture the properties of the real world. We believe that this is due to the lack of suitable dataset of annotated acoustic events from multiple microphones deployed on a car.

In the area of road-tyre interaction, we can find several studies in acoustics that analyse the properties of the wheels and conclude that road-tyre interaction generates distinctive sounds. In an autonomous driving context, such sounds could have a series of applications. Firstly, they can be used for the identification of road markers. Acoustic reports provide, indeed, encouraging results regarding the feasibility of this task; yet, the automation of road marker detection and its incorporation to the vehicle navigation system has not be done. Another interesting application is road status classification. In this space, there are works that detect the weather effect on the road (*e.g.* wetness, ice). Yet, road status classification can be extended to other types of road conditions, such as, for example, identification of damaged asphalt and potholes. A final use of road-tyre interaction sounds is terrain classification. This area is primarily explored in an off-road context and could be extended to urban scenarios (*e.g.* car parks and cobblestone streets).

Literature suggests that monitoring the ego-noise of the vehicle can be used for the timely detection of engine faults and misbehaviour. This, in turn, can be used to increase safety and reduce maintenance costs. In this space, future work could focus on the incorporation of engine fault detection in an integrated system that could make use of these anomalous sounds in real-time for navigation and planning.

In all those application areas, the case of auditory perception can be strengthened by systems that are robust to harsh noisy conditions. This is something that is rarely addressed in the current literature, limiting the applicability of auditory perception in real-world environments. Future work could focus on leveraging the advances of machine learning to develop more intelligent noise removal algorithms.

A general observation is that, besides a few notable exceptions, audio sensing is currently underrepresented in autonomous vehicles. This is largely because of the complementary nature of auditory perception. To this end, the audio

signal processing research community could investigate multi-modal frameworks that combine audio sensing with traditional sensing modalities, such as cameras and laser, to quantify the added value of auditory perception. Inspiration could be derived from the work on vehicle classification conducted in [31], which demonstrates that the audio-visual classifier significantly outperforms the visual-only classifier. For instance, future work could combine auditory odometry with traditional odometry to demonstrate the benefits of a multi-modal framework. As shown in other research areas, deep learning has great potential in this direction, as it provides architectures able to represent and learn features over multiple modalities (*e.g.* [56]). Furthermore, it offers novel effective weak supervision learning paradigms to fuse multi-modal information (*e.g.* [11]), reducing the labelling effort. An example in the autonomous driving context is given in [71], where a system for the identification of permissible driving routes from raw radar scans is weakly trained through an audio-based one that is able to predict the terrain type underneath the vehicle.

For this vision to be realised, the perception and navigation systems of the vehicle need also to adhere to strict timing constraints and, thus address the challenge of latency between sensing and prediction, due to transmission, processing and waiting delays. To this end, future systems will need to build upon theoretical foundations in the literature [43, 44, 45].

Finally, future research could focus on futuristic human-centric applications, addressing challenges in human-vehicle interaction. For example, pedestrians could hail autonomous taxis using voice communication and semi-autonomous vehicles could take control of the wheel if the driver sounds intoxicated.

Auditory perception arms autonomous vehicles with sensory capability that complements traditional sensing modalities, operating effectively in harsh light and weather conditions. In parallel, it enables new applications that enhance the safety of autonomous and semi-autonomous driving.

REFERENCES

[1] "General Motors - Self-Driving Safety Report," https://www.gm.com/content/dam/company/docs/us/en/gmcom/gmsafetyreport.pdf, 2018, *General Motors*. Online. Accessed 07.05.2019.

[2] "Din141610: Sound warning devices for authorized emergency vehicles," pp. 1–8, Jan. 2009.

[3] "International Standard IEC 61672: 2003: Electroacoustics–Sound Level Meters," 2003.

[4] "ISO 11819-1 Acoustics — Measurement of the influence of road surfaces on traffic noise — Part 1: Statistical Pass-By method," pp. 1–27, 1997.

[5] "Tesla Vehicle Safety Report," https://www.tesla.com/VehicleSafetyReport, 2019, *TESLA*. Online. Q3 2018, Q4 2018, Q1 2019. Accessed 07.05.2019.

[6] I. Abdić, L. Fridman, D. E. Brown, W. Angell, B. Reimer, E. Marchi, and B. Schuller, "Detecting road surface wetness from audio: A deep learning approach," in *23rd Int Conf Pattern Recognition (ICPR)*. IEEE, Dec. 2016, pp. 3458–3463.

[7] W. M. Adaileh, "Engine fault diagnosis using acoustic signals," in *Applied Mechanics and Materials*, vol. 295-298. Trans Tech Publ, Feb. 2013, pp. 2013–2020.

[8] R. Ahmed, M. El Sayed, S. A. Gadsden, J. Tjong, and S. Habibi, "Automotive internal-combustion-engine fault detection and classification using artificial neural network techniques," *IEEE Trans. Veh. Tech.*, vol. 64, no. 1, pp. 21–33, Apr. 2014.

[9] J. Alonso, J. López, I. Pavón, M. Recuero, C. Asensio, G. Arcas, and A. Bravo, "On-board wet road surface identification using tyre/road noise and support vector machines," *Applied acoustics*, vol. 76, pp. 407–415, Feb. 2014.

[10] M. Asayesh, B. Khodabandeloo, and A. Siami, "A random decrement technique for operational modal analysis in the presence of periodic excitations," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 223, no. 7, pp. 1525–1534, Mar. 2009.

[11] D. Barnes, W. Maddern, and I. Posner, "Find Your Own Way: Weakly-Supervised Segmentation of Path Proposals for Urban Autonomy," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, Jun. 2017. [Online]. Available: https://arxiv.org/abs/1610.01238

[12] W. Bernard and D. S. Samuel, "Adaptive signal processing," *Englewood Cliffs, NJ: Prentice Hall*, 1985.

[13] H. Bischof, M. Godec, C. Leistner, B. Rinner, and A. Starzacher, "Autonomous audio-supported learning of visual classifiers for traffic monitoring," *IEEE Intelligent Systems*, vol. 25, no. 3, pp. 15–23, Mar. 2010.

[14] C. A. Brooks and K. Iagnemma, "Vibration-based terrain classification for planetary exploration rovers," *IEEE Transactions on Robotics*, vol. 21, no. 6, pp. 1185–1191, Dec. 2005.

[15] ——, "Self-supervised terrain classification for planetary surface exploration rovers," *Journal of Field Robotics*, vol. 29, no. 3, pp. 445–468, Jan. 2012.

[16] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. IEEE, May 2006, pp. 96–100.

[17] J. Chen, S. Tian, H. Xu, R. Yue, Y. Sun, and Y. Cui, "Architecture of vehicle trajectories extraction with roadside lidar serving connected vehicles," *IEEE Access*, vol. 7, pp. 100 406–100 415, Jul. 2019.

[18] S. Dandare and S. Dudul, "Support vector machine based multiple fault detection in an automobile engine using sound signal," *Journal of Electronic and Electrical Engineering*, vol. 3, no. 1, pp. 59–63, Mar. 2012.

[19] N. Dayong, S. Changle, G. Yongjun, Z. Zengmeng, and H. Jiaoyi, "Extraction of fault component from abnormal sound in diesel engines using acoustic signals," *Mechanical Systems and Signal Processing*, vol. 75, no. 6, pp. 544–555, Dec. 2015.

[20] R. A. De Lorenzo and M. A. Eilers, "Lights and siren: A

review of emergency vehicle warning systems," *Annals of emergency medicine*, vol. 20, no. 12, pp. 1331–1335, 1991.

[21] G. Descornet, "Vehicle noise emission on wet road surfaces," in *Proc. 29th Int. Congress Noise Control Engineering*, Aug. 2000.

[22] M. Di Felice, R. Doost-Mohammady, K. R. Chowdhury, and L. Bononi, "Smart radios for smart vehicles: Cognitive vehicular networks," *IEEE Vehicular Technology Magazine*, vol. 7, no. 2, pp. 26–33, Jun. 2012.

[23] H. Ding, J. Lu, X. Qiu, and B. Xu, "An adaptive speech enhancement method for siren noise cancellation," *Applied Acoustics*, vol. 65, no. 4, pp. 385–399, Apr. 2004.

[24] B. Fazenda, H. Atmoko, F. Gu, L. Guan, and A. Ball, "Acoustic based safety emergency vehicle detection for intelligent transport systems," in *Proceedings of the IEEE ICROS-SICE International Joint Conference 2009*. IEEE, Aug. 2009, pp. 4250–4255.

[25] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International journal of computer vision*, vol. 61, no. 1, pp. 55–79, Jan. 2005.

[26] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[27] E. Freitas, P. Pereira, L. de Picado-Santos, and A. Santos, "Traffic noise changes due to water on porous and dense asphalt surfaces," *Road Materials and Pavement Design*, vol. 10, no. 3, pp. 587–607, Sep. 2009.

[28] A. Gail and W. Bartolomaeus, "Noise emission of structured road markings," *Procedia-Social and Behavioral Sciences*, vol. 48, pp. 544–552, Apr. 2012.

[29] A. Gail, W. Bartolomaeus, and M. Zoller, "Influence of surface textures of road markings on tyre/road marking noise," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 249, no. 4, Sep. 2014, pp. 3473–3482.

[30] W. Gardziejczyk, "Comparison of vehicle noise on dry and wet road surfaces," *Foundation of Civil and Environmental Engineering*, vol. 9, pp. 5–15, Jan. 2007.

[31] M. Godec, C. Leistner, H. Bischof, A. Starzacher, and B. Rinner, "Audio-visual co-training for vehicle classification," in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, Aug. 2010, pp. 586–592.

[32] H. Grabner and H. Bischof, "On-line boosting and vision," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Jun. 2006, pp. 260–267.

[33] C. Gurau, D. Rao, C. H. Tong, and I. Posner, "Learn from experience: Probabilistic prediction of perception performance to avoid failure," *The International Journal of Robotics Research*, vol. 37, no. 9, pp. 981–995, Oct. 2017.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[35] S. Ishida, S. Liu, K. Mimura, S. Tagashira, and A. Fukuda, "Design of acoustic vehicle count system using DTW," in *Proc. ITS World Congress*, Oct. 2016, pp. 1–10.

[36] S. Ishida, J. Kajimura, M. Uchino, S. Tagashira, and A. Fukuda, "Saved: Acoustic vehicle detector with speed estimation capable of sequential vehicle detection," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Nov. 2018, pp. 906–912.

[37] M. Kemmler, E. Rodner, E.-S. Wacker, and J. Denzler, "One-class classification with gaussian processes," *Pattern Recognition*, vol. 46, no. 12, pp. 3507–3518, Dec. 2013.

[38] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[39] W. Kongrattanaprasert, H. Nomura, T. Kamakura, and K. Ueda, "Detection of road surface states from tire noise using neural network analysis," *IEEJ Transactions on Industry Applications*, vol. 130, no. 7, pp. 920–925, Aug. 2010.

[40] W. A. Leasure Jr and E. K. Bender, "Tire–road interaction noise," *The Journal of the Acoustical Society of America*, vol. 58, no. 1, pp. 39–50, Jul. 1975.

[41] C. Leistner, A. Saffari, P. M. Roth, and H. Bischof, "On robustness of on-line boosting-a competitive study," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE, Sep. 2009, pp. 1362–1369.

[42] R. Lenain, B. Thuilot, C. Cariou, and P. Martinet, "High accuracy path tracking for vehicles in presence of sliding: Application to farm vehicle automatic guidance for agricultural tasks," *Autonomous robots*, vol. 21, no. 1, pp. 79–97, Jun. 2006.

[43] Z. Li, C. Huang, and H. Yan, "Stability analysis for systems with time delays via new integral inequalities," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 12, pp. 2495–2501, 2018.

[44] Z. Li, H. Yan, H. Zhang, X. Zhan, and C. Huang, "Improved inequality-based functions approach for stability analysis of time delay system," *Automatica*, vol. 108, p. 108416, 2019.

[45] Z. Li, H. Yan, H. Zhang, Y. Peng, J. H. Park, and Y. He, "Stability analysis of linear systems with time-varying delay via intermediate polynomial-based functions," *Automatica*, vol. 113, p. 108756, 2020.

[46] Z. Li, S. Akishita, and T. Kato, "Engine failure diagnosis with sound signal using wavelet transform," SAE Technical Paper, Tech. Rep., Feb. 1997.

[47] J. Libby and A. J. Stentz, "Using sound to classify vehicle-terrain interactions in outdoor environments," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, May 2012, pp. 3559–3566.

[48] R. F. Lyon, A. G. Katsiamis, and E. M. Drakakis, "History and future of auditory filter models," in *Proceedings of 2010 IEEE International Symposium on Circuits and*

Systems. IEEE, May 2010, pp. 3809–3812.

[49] M. Madain, A. Al-Mosaiden, and M. Al-khassaweneh, "Fault diagnosis in vehicle engines using sound recognition techniques," in *2010 IEEE International Conference on Electro/Information Technology*, 2010, pp. 1–4.

[50] L. Marchegiani and P. Newman, "Learning to listen to your ego(-motion) : Metric motion estimation from auditory signals," in *Towards Autonomous Robotics Systems (TAROS)*, Jul. 2018, pp. 247–259.

[51] ——, "Listening for sirens: Locating and classifying acoustic alarms in city scenes," *arXiv:1810.04989*, Oct. 2018.

[52] L. Marchegiani and I. Posner, "Leveraging the urban soundscape: Auditory perception for smart vehicles," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2017, pp. 6547–6554.

[53] F. Meucci, L. Pierucci, E. Del Re, L. Lastrucci, and P. Desii, "A real-time siren detector to improve safety of guide in traffic environment," in *16th European Signal Processing Conference*, 2008, pp. 1–5.

[54] D. Naish, "A study on the sound power of queensland road vehicles," in *Proc. 20th Int. Congress on Acoustics*, Aug. 2010, pp. 1–8.

[55] M. K. Nandwana and T. Hasan, "Towards smart-cars that can listen: Abnormal acoustic event detection on the road." in *INTERSPEECH*, Sep. 2016, pp. 2968–2971.

[56] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, Jun. 2011, pp. 689–696.

[57] S. Odedra, S. D. Prior, M. Karamanoglu, M. A. Erbil, and S.-T. Shen, "Using acoustic sensor technologies to create a more terrain capable unmanned ground vehicle," in *Int. Conf. Eng. Psychology and Cognitive Ergonomics*. Springer, Jul. 2009, pp. 574–579.

[58] A. Pico, G. Schillaci, V. V. Hafner, and B. Lara, "How do I sound like? forward models for robot ego-noise prediction," in *IEEE Int Conf Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, Sep. 2016, pp. 246–251.

[59] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[60] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Medical image computing and computer-assisted intervention*. Springer, Oct. 2015, pp. 234–241.

[61] U. Sandberg, "Road traffic noise—the influence of the road surface and its characterization," *Applied Acoustics*, vol. 21, no. 2, pp. 97–118, 1987.

[62] J. Schröder, S. Goetze, V. Grutzmacher, and J. Anemüller, "Automatic acoustic siren detection in traffic noise by part-based models." in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2013, pp. 493–497.

[63] J. Stewart, "Driverless Cars Need Ears as well as Eyes," https://www.wired.com/story/driverless-cars-need-ears-as-well-as-eyes/, Aug. 2017, *WIRED*. Online. Accessed 07.05.2019.

[64] A. Valada and W. Burgard, "Deep spatiotemporal models for robust proprioceptive terrain classification," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1521–1539, Aug. 2017.

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

[66] R. P. Wagner, "Guide to test methods, performance requirements, and installation practices for electronic sirens used on law enforcement vehicles," NIST, Tech. Rep., Aug. 2000.

[67] M. M. Waldrop, "Autonomous vehicles: No drivers required," *Nature News*, vol. 518, no. 7537, p. 20, 2015.

[68] J. Walker, "The Self-Driving Car Timeline – Predictions from the Top 11 Global Automakers," https://emerj.com/ai-adoption-timelines/self-driving-car-timeline-themselves-top-11-automakers/, Mar. 2020, *EMERJ*. Online. Accessed 22.05.2020.

[69] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.

[70] Y. Wang, Y. Xing, and H. He, "An intelligent approach for engine fault diagnosis based on wavelet preprocessing neural network model," in *The 2010 IEEE International Conference on Information and Automation*. IEEE, Jun. 2010, pp. 576–581.

[71] D. Williams, D. De Martini, M. Gadd, L. Marchegiani, and P. Newman, "Keep off the grass: Permissible driving routes from radar with weak audio supervision," in *The 23rd IEEE International Conference on Intelligent Transportation Systems*. IEEE, Sep. 2020.

Letizia Marchegiani received a PhD in Computer Engineering from Sapienza - University of Rome (Italy) in 2012. She is currently an Assistant Professor at Aalborg University (Denmark). Her research interests are in the areas of signal processing, machine learning, and their application to robotics, autonomous systems, cognitive modelling and intelligent transportation. Postal Address: Fredrik Bajers Vej 7, 9220 Aalborg Ø, Denmark. Email: lm@es.aau.dk

Xenofon Fafoutis received a PhD in Embedded Systems Engineering from the Technical University of Denmark (DTU) in 2014. He is currently an Associate Professor at the same university. His research interests are in Wireless Embedded Systems and in emerging applications, such as Intelligent Transportation and the Industrial Internet of Things. Postal Address: Richard Petersens Plads, Building 322, 2800 Kgs. Lyngby, Denmark. Email: xefa@dtu.dk