Leveraging the Urban Soundscape: Auditory Perception for Smart Vehicles

Letizia Marchegiani[1] and Ingmar Posner[1]

***Abstract*— Urban environments are characterised by the presence of distinctive audio signals which alert the drivers to events that require prompt action. The detection and interpretation of these signals would be highly beneficial for smart vehicle systems, as it would provide them with complementary information to navigate safely in the environment. In this paper, we present a framework that spots the presence of acoustic events, such as horns and sirens, using a two-stage approach. We first model the urban soundscape and use anomaly detection to identify the presence of an anomalous sound, and later determine the nature of this sound. As the audio samples are affected by copious non-stationary and unstructured noise, which can degrade classification performance, we propose a noise-removal technique to obtain a clean representation of the data we can use for classification and waveform reconstruction. The method is based on the idea of analysing the spectrograms of the incoming signals as images and applying spectrogram segmentation to isolate and extract the alerting signals from the background noise. We evaluate our framework on four hours of urban sounds collected driving around urban Oxford on different kinds of road and in different traffic conditions. When compared to traditional feature representations, such as Mel-frequency cepstrum coefficients, our framework shows an improvement of up to 31% in the classification rate.**

I. INTRODUCTION

Smart vehicle systems offer a unique opportunity towards the realisation of applications and services of high socioeconomic impact that can revolutionise everyday life, offering a safer and more comfortable means of transportation. Auditory perception and sound processing can play a crucial role in this context. In fact, in a driving scenario, certain alerting stimuli, such as sirens, are meant to be heard, and some of them, such as horns, are exclusively acoustic. The majority of research in robotics and smart autonomous systems has focused on optical sensors, radars and lasers as means of interpreting the environment. In several situations, however, acoustic signals provide complementary information that cannot be captured by traditional sensing modalities. Indeed, hearing enables omnidirectional perception and overcomes the limitations imposed by occlusions. Various sounds can be used as cues for events that require further attention and, as a result, help avoid danger by navigating the focus on things that require prompt action. It is apparent that people with hearing impairments are potentially more prone to accidents that could be avoided if such cues could be perceived [1]. In the same way, autonomous cars would highly benefit from the ability to identify and interpret acoustic signals which carry crucial information in traffic scenarios. An emergency vehicle approaching an intersection could be detectable much before it reaches the crossing point and despite potential occlusions. The possibility of having advance information of this kind would considerably increase the time frame allowed for a safe response from the driver and, for a smart car working in semi-autonomous regime could also be used to trigger manual intervention.

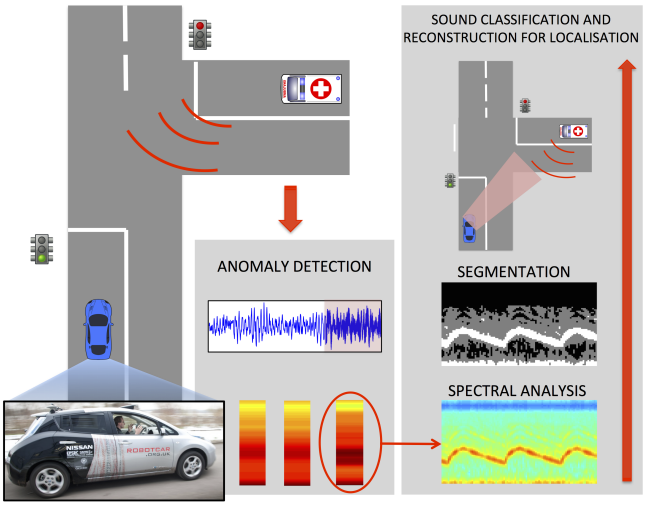


Fig. 1. Example of a typical use-case with an emergency vehicle approaching an intersection and the proposed pipeline. The framework first detects the presence of an anomalous sound and then further process this sound to classify the correspondent sound source and reconstruct the waveform. Such processing includes spectral analysis through the use of Gammatonegram representation and spectrogram segmentation to extract a binary mask used for classification, target signal extraction and waveform reconstruction. The reconstructed signal in the time domain allows the localisation of the sound source with off-the-shelf algorithms.

In this paper we leverage auditory perception in smart vehicles for urban soundscape modelling and interpretation. We propose a framework to detect acoustic events such as sirens, horns, and pedestrian traffic lights (i.e. accessible traffic signals). Rather than directly targeting specific sound events and employing filtering techniques to spot the presence of those sounds, we model the urban road traffic soundscape and we use anomaly detection to determine the presence of an acoustic event. As the considered sounds are conceived to overcome background noise [2], their temporal and spectral characteristics are meant to differ from the ones of the background noise to be easily audible and attract the attention of the driver (see e.g. [3] and [4]). This allows us to identify them through anomaly detection. Moreover, such a modelling choice enables the framework to detect acoustic events which are neither sirens, horns or pedestrian traffic lights, but that still represent an anomaly and need to be

[1]Authors are members of the Oxford Robotics Institute, University of Oxford, United Kingdom, {letizia, ingmar}@robots.ox.ac.uk

considered.

Our analysis is twofold: we first detect the presence of anomalous sounds using one-class classification and then further process the detected anomalous sounds to identify their nature. One of the main challenges of this process lies in the presence of non-stationary and unstructured noise in the data. The audio samples, in fact, are mixtures of an unknown number of different static and dynamic sound sources, whose characteristics and spatial distribution are not available. Classic feature representations used in the literature, such as Mel-frequency cepstrum coefficients [5], provide a compact and efficient representation of the shape of the spectrum of a signal, and have been shown to perform well on clean data. However, their discriminative power decreases when dealing with more realistic and complex scenarios. Moreover, acoustic events, such as the ones analysed in this paper, can have a highly variable duration both within the same class of sound (e.g. horns) and between different classes (e.g. sirens and horns). This highlights the importance of considering the temporal dynamics of the signals to model the urban soundscape.

With this purpose, we opt for spectrogram-based representations of the sounds, as they incorporate information both in the time and the frequency domains, and provide a visual signature of the signals. We take advantage of such a visual signature, proposing a noise-removal method, which relies on the idea of treating the spectrograms as one-channel images and applying image processing techniques. Specifically, we perform spectrogram segmentation to extract the time-frequency content of the target signal from the one of the background noise. This process allows us to perform classification on a time-frequency representation of the sound samples which is no longer affected by the presence of maskers, overcoming the limitations induced by the high variability of the noise to produce more reliable results. Finally, as the time-frequency content of the target signal has been identified and isolated, the original waveform in the time domain is reconstructed and can potentially be used to localise the respective sound source. A representation of the entire framework is provided in Figure 1. We evaluate our framework on four hours of urban acoustic data collected driving around urban Oxford on different kinds of road and at different times of the day (i.e. different traffic conditions). To the best of our knowledge, this is the first work investigating model-based anomalous acoustic event detection, classification and waveform reconstruction in driving scenarios for smart car applications. It is also the first work utilising spectrogram segmentation for signal isolation and extraction in outdoor unstructured acoustic environments.

II. RELATED WORK

Abnormal sound detection methods have already been proposed in the literature, such as in [6] and [7]. The former presents a framework for detection of abnormal sound events in a subway station by learning a set of mixtures of temporal trajectories and spotting the events that differ from the set of learnt trajectories. In the latter, the authors are able to classify abnormal sounds such as gun shots, glass breaking, and explosions in indoor environments (e.g. a rental apartment) using Hidden Markov Models (HMMs) [8]. Siren and, more generally, emergency vehicle detection in synthetic and realistic driving scenarios has been investigated in [9],[10], and [11]. In [9] and [11], the analysis relies on the knowledge of the specific frequencies characterising the pitch of the siren sound and classic signal processing filtering techniques for the detection of those frequencies. Specifically, the first work employs a two-state scheme, based on Module Difference Function (MDF) and peak searching. The second work makes use of an *adaptive predictor noise canceller* system [12] to identify and extract the siren signal. In [10] a part-based model is proposed and its performance analysed in a simulated driving scenario at different signal-to-noise ratio (SNR) levels. The authors obtain good results in clean scenarios, but the performance decreases with higher levels of noise. In robotics, much interest has been devoted to speech processing for human-robot interaction purposes (e.g. [13]), while non-speech sound modelling has received less investigation. Indoor acoustic event classification for domestic robot applications has been explored in [14] and articulated in two different phases: a preliminary sound detection phase and a final sound recognition one. Additional acoustic event recognition has been proposed in [15] and [16]. In these works, the audio data is modelled using traditional feature selections in the time and in the frequency domains, with the Mel-frequency cepstrum coefficients (MFCCs) [5] being the most commonly employed.

This paper is in line with the two-phase structure presented in [14], but, rather than using cross-correlation (which emphasises the changes in the temporal evolution of the signal) as a measure of difference to identify anomalous signals, it models the driving soundscape and defines anomalous sounds as the ones which are less likely to be encountered in such a soundscape. This allows us to obtain a long-term representation of the acoustic environment, which takes into account different noisy scenarios with variegated time-frequency patterns, leading to a more robust anomaly detection. Such variety cannot always be captured by cross-correlation similarity metrics. In this fashion, we share the aspiration of [6]. With respect to that work, though, we also provide a classification scheme to assign the detected anomalous sounds to specific groups of interest, such as sirens, pedestrian traffic lights and horns. Furthermore, as we are considering environments characterised by unstructured and non-stationary noise, we cannot rely on more traditional feature representations, which are sensitive to different levels of auditory masking. With this purpose, we consider the visual signatures of the signals drawn by their spectrograms. Works in other disciplines have used the visual insights of the spectrogram as input for further analysis (e.g. [17][18]). Sharing this vision, in this paper we treat the spectrograms as one-channel images, applying vision-based segmentation techniques to isolate the anomalous signal from the background noise in a complete unsupervised way. Such representation allows us to overcome the limitations imposed

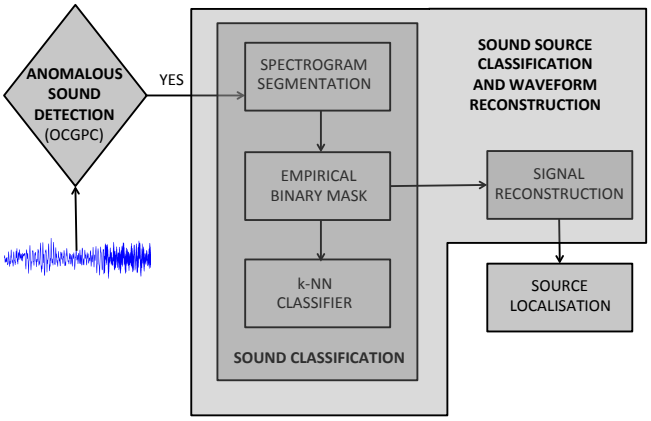


Fig. 2. Flowchart of the pipeline operating in the framework. Two main phases can be identified: anomalous sound detection, and sound source classification and waveform reconstruction. The first phase is based on One-Class Gaussian Process Classification (OCGPC). When anomalous sounds are detected at this stage, further processing of those sounds is performed. The spectrogram representation of the signal is segmented to extract the target sound from the background noise, generating Empirical Binary Masks (EBMs). Those masks are fed to a k-NN classifier, which identifies the nature of the sound source. The same masks are used, by comparison with the original spectrogram of the noisy signals (target together with the background noise) to reconstruct the target signal, which can be used as input to off-the-shelf localisation algorithms.

by the presence of unstructured and non-stationary noise in the audio data, leading to a more accurate and robust classification. Furthermore, this feature selection offers also the inherent possibility of extracting the target signal from the background, which can be used for reconstruction and sound source localisation.

III. TECHNICAL APPROACH

As already proposed by other works in the field (e.g. [13], [14]), our framework is articulated in two different phases. Figure 2 shows in more detail the pipeline operating in the framework. Following a two-phase approach is computationally advantageous, as, in this way, only the anomalous sounds are further processed. Considering the system working in real-time on a continuous audio input stream, the anomalous sounds will represent a small portion of this stream, hence the high computational savings.

A. Anomalous Sound Detection

Several approaches have been proposed in the literature to solve anomaly detection tasks, following a supervised, unsupervised or semi-supervised paradigm. In our framework, we employ a one-class classification approach, as it is easy to obtain data from the *normal* class, while the availability of samples from the *anomalous* class is quite limited and such samples are not necessarily representative of the entire class. In particular, we implement One-Class Gaussian Process Classification (OCGPC). With respect to other kernel-based methods, OCGPC provides a Bayesian framework and it has shown to outperform other techniques, such as Support Vector Data Description [19] in several one-class classification tasks [20]. OCGPC is a special case of Gaussian Process (GP) binary classification when only samples from one of two classes are provided in the training phase. In the case of standard GP binary classification, given a training set $\mathcal{X} = \{\mathbf{x_1}, \mathbf{x_2}, .., \mathbf{x_n}\}$ of n examples $\mathbf{x_i} \in \mathbb{R}^D$, denoting feature vectors and corresponding binary labels $y_i \in \{-1, 1\}$, the goal is to predict the label y^* of an unseen example $\mathbf{x}^*$. Classification is obtained by introducing a latent function $f(\mathbf{x})$, and then applying any sigmoidal function $\sigma(f)$ such that $p(y^* = 1|\mathbf{x}^*) = \sigma(f(\mathbf{x}^*))$, to squash the output of the latent function into $[0, 1]$. The prior on the latent function can be modelled as a GP, $f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$, specified by a mean function $\mu(\mathbf{x})$ and a covariance (or kernel) function $k(\mathbf{x}, \mathbf{x}')$. Given a specific covariance function, the values of the hyper-parameters are obtained during the training phase, by maximising the log marginal likelihood of the training data. Predictions on an unseen example $\mathbf{x}^*$ are obtained in two steps. In the first one, the distribution over the latent variable corresponding to the unseen example is obtained using:

$$p(f^*|\mathcal{X}, \mathbf{y}, \mathbf{x}^*) = \int p(f^*|\mathcal{X}, \mathbf{x}^*, \mathbf{f}) p(\mathbf{f}|\mathcal{X}, \mathbf{y}) \, \mathrm{d}\mathbf{f}. \quad (1)$$

In the second one, the probabilistic prediction is obtained by marginalisation of the latent function $f^* = f(\mathbf{x}^*)$:

$$p(y^* = 1|\mathcal{X}, \mathbf{y}, \mathbf{x}^*) = \int \sigma(f^*) p(f^*|\mathcal{X}, \mathbf{y}, \mathbf{x}^*) \, \mathrm{d}f^*. \quad (2)$$

Exact inference is not possible due to the sigmoidal function, but approximations, such as Laplace approximation (LA) and Expectation Propagation (EP) [21] can be used. In OCGPC, only the samples from one class will be provided at training phase. In this case, only feature vectors $\mathbf{x}$ correspondent to normal (i.e. non anomalous) audio frames ($y = 1$) will be used for training. However, it is possible to derive membership scores by choosing a specific GP prior characterised by a mean function with a smaller value than the one used to indicate the positive class labels [20]. Specifically, we choose the prior on the latent function to be characterised by a zero mean function and covariance function K. This choice will reduce the space of probable latent functions to functions with values gradually decreasing while far away from observed points and give the possibility of using directly the predictive probability $p(y^* = 1|\mathcal{X}, \mathbf{y}, \mathbf{x}^*)$ and its first and second order moments as membership scores. The posterior mean function, in fact, will have high values (around $y = 1$) in high density areas close to the training points and monotonically decreasing values while far away from the observed points. The posterior variance will be characterised, instead, by the opposite behaviour. For further details and for a mathematical justification on predictive mean and variance being suitable OCC measures, the reader is referred to [20].

B. Sound Classification and Waveform Reconstruction

When an incoming signal $\mathbf{x}^*$ is considered anomalous, further processing is applied. Specifically, we use a spectrogram-based representation of the audio sample, car-

rying information both in the time domain and frequency domains. As we are dealing with realistic scenarios, such representation, however, is corrupted by the presence of noise, whose characteristics are not known a priori, and that can highly affect classification performance. Taking advantage of the visual appearance of the spectrogram, where masker and target sounds are easily identifiable, we treat the spectrogram of the signal as a one-channel image and employ image segmentation techniques to perform noise removal. Specifically, we apply k-means image segmentation [22], as it operates in a completely unsupervised manner and without the necessity of specific input knowledge on the structure of the specific class of the segment. The segmentation allows us to cluster different portions of the spectrogram depending on their energy and to identify and isolate the cluster characterised by the highest energy content. In this case, given the nature of the considered anomalous sounds, the highest energy content corresponds to the target signal (i.e. anomalous sound without background noise) [2]. Let us define as $\mathcal{C} = \{C_1, C_2, ..., C_k\}$ the set of clusters obtained after k-means segmentation where k is the total number of clusters, and as C_{max} the cluster containing the most powerful parts of the spectrogram. Empirical Binary Masks (EBMs) can be obtained by assigning value 1 to all the time-frequency bins of the spectrogram which are part of C_{max}, while setting to 0 the time-frequency bins in $\mathcal{C}\backslash\{C_{max}\}$. EBMs are conceptually akin to Ideal Binary Masks (IBM) [23]. Both representations, in fact, aim to identify the most powerful time-frequency bins of a spectrogram (corresponding the a specific target signal) with respect to an interference one. More details on the characteristics of IBMs and their use as a solution to Computational Acoustic Scene Analysis (CASA) are provided in [23], [24] and [25].

EBMs are then fed to a k-NN based classifier [26] to identify the nature of the sound source. The k-NN classifier is trained using samples from the different classes that we are interested in recognising, and classifies the anomalous audio frames in input by a majority vote of its k nearest neighbours. It has been shown that it is possible to reconstruct (re-synthesise) the waveform of the estimated target signal by combining information from the gammatonegram of the original mixture signal (target signal together with background noise) and the binary mask related to the target signal alone [27][28]. When more than one channel (microphone) is available, sound source localisation can be achieved using interaural time difference and cross-correlation, as illustrated in [29].

IV. EXPERIMENTAL EVALUATION

To evaluate the performance of our framework, we collected four hours of data by driving around Oxford, UK. The data has been gathered using two Knowles omnidirectional boom microphones mounted on the roof of a car and an ALESIS IO4 audio interface. The data has been recorded at a sampling frequency f_s of 44100 Hz at a resolution of 16 bits. Additional data used to train the k-NN classifier has been taken from the Urban Sound Dataset [30]. A more detailed description of the dataset employed is given in Table I.

	Training			Testing		
NR	320,000			17,000		
AN	Siren	Horn	PTL	Siren	Horn	PTL
	3496	3168	3210	106	98	93

TABLE I. Dataset used for evaluation. The table shows the distribution of the samples used for training and testing. NR denotes the normal audio samples, while AN indicates the anomalous ones, which are divided into siren, horn and pedestrian traffic light (PTL). Samples here refer to frames of one second used for detection and classification.

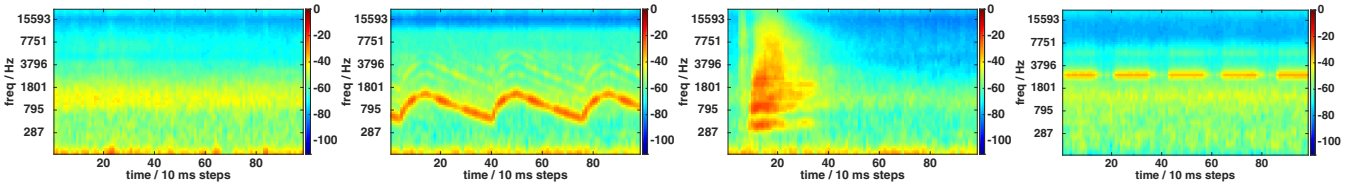
A. Feature Representation

Previous studies (e.g. [6]) have shown that for sound event detection and classification tasks, standard features in the frequency domain are not guaranteed to provide satisfactory performance. In realistic and noisy scenarios, in fact, and while dealing with classes of signals of variable duration, information about the frequency components of the audio signals, as well as their evolution in the time-domain, need to be taken into account.

Classic spectrograms are a visual representation of the Short-Time Fourier Transform (STFT) of a signal. In this paper we use a special case of spectrograms, the *gammatonegrams*, which are characterised by an STFT obtained with gammatone filterbanks [31]. Gammatone filterbanks have been originally introduced in [32] as an approximation of the human cochlea signal analysis and, for this reason, are called *perceptual features*. As the specific sounds we are considering are meant to differ from the masking noise, to be easily detectable by the drivers, we prefer using a representation able to mimic the behaviour of the human auditory system and, thus, to capture these differences. Moreover, gammatonegrams have been shown to be more suitable than other time-frequency representations for waveform reconstruction [33]. Standard spectrograms are obtained by considering a constant bandwidth across all frequency channels. In the spectral representation which derives from the application of gammatone filterbanks, instead, the frequency bins are not equally spaced. Gammatonegrams have been proved to be extremely useful features in several audio classification tasks (e.g. [34]). Figure 3 shows an example of gammatonegrams for the considered acoustic classes. Specifically, we use 64 frequency channels between 50 Hz and $f_s/2 = 22050$ Hz, corresponding to the maximum frequency resolution given by the STFT. The gammatonegrams are computed on time domain frames of one second after applying a Hamming window to avoid spectral leakage.

B. Anamalous Sound Detection

Following [20], we train an OCGPC using only normal sounds. The detection is then performed by using both normal and anomalous sounds. Since the frequency content characterising the acoustic signals belonging to the two classes



(a) Normal (b) Siren (c) Horn (d) Pedestrian Traffic Light

Fig. 3. Example of the Gammatonegram representation of sound frames of one second for the different acoustic classes considered. From left to right: normal sound (i.e. non anomalous), siren, horn, and pedestrian traffic light. The energy of the time-frequency bins is expressed in decibel (dB) scale. We observe that the frequency bins are not equally spaced, due to the application of the gammatone filterbanks.

(normal and anomalous) differs significantly, we used a time-independent version of the gammatonegrams, obtained by first summing the time-frequency bins over time and then normalising the power values of the resulting frequency bins. Figure 4 shows the time-independent representation for the gammatonegrams illustrated in Figure 3.

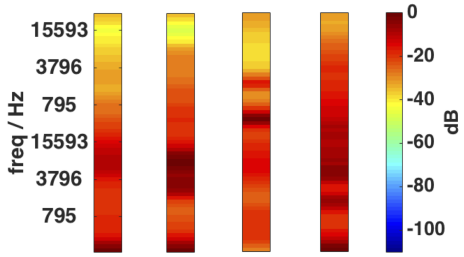


Fig. 4. Time-independent and normalised representation of the gammatonegrams in Figure 3. From left to right: normal sound, siren, horn, and pedestrian traffic light.

Figure 5 shows the detection performance when a radial basis function (RBF) kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}'\|^2)$ is used. The figure provides the Receiver Operating Characteristic (ROC) curve both in case of decisions based on the GP posterior mean μ^* and the GP posterior variance σ^{*2}. The area under the curves (AUC) is given in the legend. We observe that the output of the first stage constitutes a very trustworthy input for the second stage.

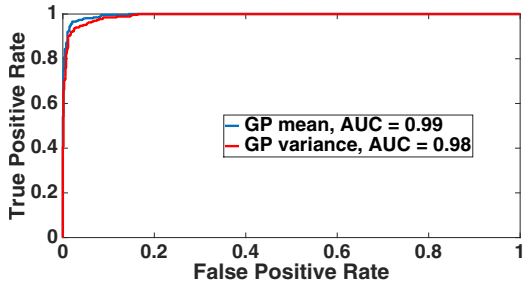


Fig. 5. ROC curves in case of decisions based on the GP posterior mean μ^* and the GP posterior variance σ^{*2}. The performance is very similar in both circumstances.

C. Sound Source Classification

The sound frames which have been considered anomalous from the previous step of the pipeline, are then classified and the respective waveform reconstructed. The classification is performed using a k-NN based classifier operating on the EBMs of the anomalous sound frames. The EBMs employed for training are obtained from a combination of samples from the Urban Sound Dataset (USD) [30], samples from our data collection with the car, and samples publicly available from `www.freesound.org` (FreeS). In particular, we are interested in three different kinds of anomalous sounds: sirens, horns and pedestrian traffic lights. We use data from the USD for what regards sirens and horns, and a combination of samples from our data collection and from FreeS for what regards the pedestrian traffic lights. Training data has been augmented through affine transformations of the original signals, to take into account potentially different scenarios at testing phase.

1) Empirical Binary Masks and Classification: The first step towards EBMs is to segment the gammatonegram of the sound frames using k-means. We explore the behaviour of the segmentation considering a different number k of clusters, finding the best performance in case of $k = 3$. Once the segmentation is completed, we select the cluster containing the most powerful parts of the gammatonegram. The EBM is then obtained by setting all of the time-frequency bins in this cluster to 1 and all the remaining bins to 0. We observe that EBMs computed in this way are still characterised by some low-frequency noise, which we attribute to the presence of some normal driving noise in the background. As we follow a model-based approach for the detection of anomalous sounds, we now have the possibility to extrapolate statistics on the normal sounds, and to remove most part of the noise still present in the masks, by comparing the EBMs of the anomalous signals to the EBM extracted from the average gammatonegram of the normal sound frames that we used to train the OCGPC in Section IV-B. Specifically, we want to keep only the time-frequency regions in the target mask which are not in the binary mask of the background noise. Let us define the EBM resulting from noise removal as EBM_{clean}, the original EBM of the anomalous sound considered as EBM_{target} and the EBM related to the average gammatonegram of normal sounds as EBM_{noise}. Leveraging their binary nature, the noise removal is performed with the following binary operation: $EBM_{clean} = EBM_{target} \wedge \neg EBM_{noise}$. An example of how the clean EBM is generated is given in Figure 6. The original gammatonegram is given

in Figure 3c. Table II reports the confusion matrix obtained by averaging across different numbers of neighbours $k \in \{1, 5, 10, 15, 25, 30, 35, 40, 45, 50, 55, 100\}$. The average classification rate for all classes is shown along the diagonal of the matrix. The first column (EBMs) of Table III reports the average classification performance for different numbers of neighbours k, across the three classes.

	Predicted Class		
Class	Siren	Horn	PTL
Siren	**0.81**	0.19	0
Horn	0.18	**0.79**	0.03
PTL	0.11	0.05	**0.83**

TABLE II. Confusion Matrix obtained averaging across the different number of k. The average classification rate for all classes is shown along the diagonal of the matrix.

2) Baseline Evaluation Methods: Our feature selection inherently offers the facility for performing sound source segregation and recovering the waveform of the target signal. To further highlight the benefits provided by the use of the EBMs, we also analyse other feature representations commonly used in the literature for sound recognition tasks and compare the classification performance obtained. In particular, we compare our approach against the full gammatonegrams, computed as illustrated in Section IV-A, and the Mel-frequency cepstrum coefficients (MFCCs). MFCCs are an approximation of the envelope of the short-term power spectrum of a sound and are the most widely used representation in acoustic event detection tasks [6]. In this case, we use the first 13 coefficients together with their first and second derivatives. The results obtained for different numbers of neighbours k are shown in Table III. We can see that the use of EBMs improves overall classification with respect to the other feature selections. Table IV reports the confusion matrix (averaged across different values of k) for all classes, using the full gammatonegrams (GTGs) and the MFCCs. We can observe that the EBMs provide good performance for all the classes considered, while the GTGs and the MFCCs demonstrate imbalanced performance: they yield good results with pedestrian traffic lights, but they are less reliable with regards to horns and sirens. We attribute this difference to the fact that the pedestrian traffic light samples are generally characterised by less background noise (when the traffic light beeps, cars are generally not moving) compared to sirens and horns. This suggests that, while GTGs and MFCCs are sensitive to noise, the specific nature of the EBMs makes them more suitable for acoustic event classification in realistic noisy scenarios. Furthermore, these results highlight the necessity and the accuracy of the noise-removal technique presented in the paper and foresees the benefits of using the resulting representation for sound event detection and classification in noisy environments.

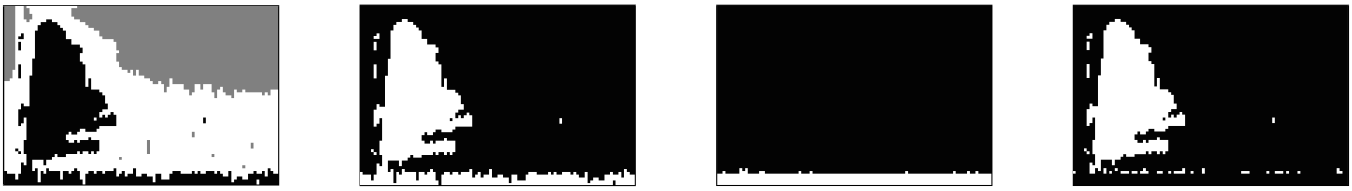
3) Unknown Anomalous Sounds: As we do not expect that every possible anomalous sound can be considered *a priori*, we also analyse the ability of our framework to recognise as unknown the anomalous sounds that are different from any sound provided in the training phase. Specifically, we use randomly chosen samples from other two classes available in the USD: street music and drilling. Several techniques have been proposed in the literature to handle these cases in a k-NN classification framework, such as [35] and [36]. We compare three different techniques: Mean Distance Factor (MDF) (described in [37]), k-th Distance Factor (KDF) (proposed in [38]) and Class Dependent Distance Factor (CDDF). The latter one is an adaptation of common clustering-based anomaly detection techniques (e.g. [39]) to a k-NN context. In case of MDF, a sample is considered to be an outlier (it does not belong to any of the classes we used to train the k-NN classifier) when the average distance of its k neighbours is higher than a given threshold. In case of KDF, a sample is considered to be an outlier when the distance from its k^th neighbour is higher than a given threshold. Both these approaches do not take into consideration the density of the different classes (*k-NN Global Anomaly Detection*). In the case of the CDDF, instead, we first compute the centroid for each of the classes in the training set. We then define as $D_i^{train}, i \in \{1, 2, 3\}$ the average distance between the centroid of class i and the training samples of the same class. For each sample in the testing set, we compute the

k	EBMs (ours)	GTG	MFCC
1	**0.82**	0.77	0.67
5	**0.80**	0.76	0.62
10	**0.81**	0.76	0.61
25	**0.83**	0.75	0.60
40	**0.83**	0.74	0.61
Average	**81.8 ± 1.30**	75.6 ± 1.14	62.2 ± 2.77

TABLE III. Average classification performance (across all classes) varying the number of neighbours k and using different feature representations: the Empirical Binary Masks proposed in this paper, the full Gammatonegrams (GTG) and the Mel-frequency cepstrum coefficients (MFCCs).

	Predicted Class					
	GTG			MFCC		
Class	Siren	Horn	PTL	Siren	Horn	PTL
Siren	**0.72**	0.28	0	**0.69**	0.06	0.25
Horn	0.37	**0.55**	0.07	0.31	**0.21**	0.48
PTL	0	0.1	**0.99**	0.04	0	**0.96**

TABLE IV. Confusion Matrices (averaging across the different number of neighbours k) obtained using different feature representations: the full Gammatonegrams (GTG) and the Mel-frequency cepstrum coefficients (MFCCs). The average classification rate for all classes is shown along the diagonal of the matrix.



(a) Segmentation (b) EBM_{target} (c) EBM_{noise} (d) EBM_{clean}

Fig. 6. Different steps to build the Empirical Binary Masks from the sound frames to use as feature representation for classification. The figure shows the case of a sound frame related to a horn. The original gammatonegram (Figure 3c) is treated as one-channel image and k-means segmentation is applied (a). The cluster containing the most powerful time-frequency bins is kept (b) and compared with the EBM related to the average gammatonegram of normal sounds (c) for additional noise removal. Final EBM used in the k-NN classifier is shown in (d).

distance D_i^{test} from the centroid of the class i it has been assigned to, according to the k-NN classifier. The sample is then considered an outlier when the ratio D_i^{test}/D_i^{train} is higher than a certain threshold. In all these methods, we use the Hamming distance. The performance obtained with the three algorithms, in case of $k = 40$ is shown in Figure 7. The figure provides the ROC curves obtained employing the MDF, KDF and CDDF. The area under the curves (AUC) is given in the legend.

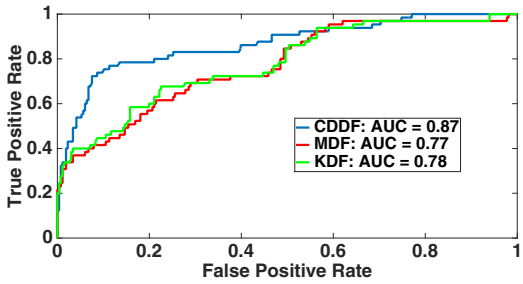
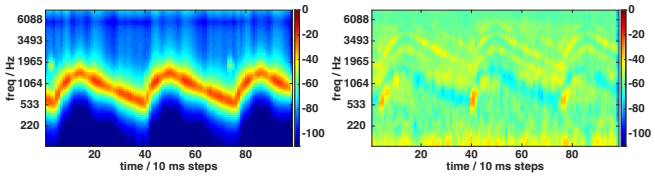


Fig. 7. ROC curve obtained applying the Mean Distance Factor (MDF) [37], the $k-th$ Distance Factor (KDF) [38] and the Class Dependent Distance Factor (CDDF) methods. The legend reports the AUC for all three cases. Best viewed in colour.

D. Waveform Reconstruction for Sound Source Localisation

When the gammatonegram of the mixture signal (target signal and background noise) and a binary mask able to separate the two are available, it is possible to re-synthesise the waveform of the target signal. The idea is to reconstruct the waveform by inverting the gammatonegram representation of the mixture signal, using the binary mask to weight the contribution of each time-frequency bin. Following [24], we first remove any across-channel phase difference. Then the phase-corrected output from each filter channel is divided into time frames, and the resulting time-frequency bins are multiplied by the correspondent weight in the mask. The re-synthesised waveform is obtained by summing the weighted filter outputs across all channels of the filterbank. Several works (e.g. [24] and [33]) have demonstrated that sound source separation using binary masks and following signal reconstruction through gammatonegram inversion provides accurate results. Quantitative evaluation is generally carried out by comparing the signal-to-noise ratio (SNR) before and after the segregation procedure. However, this requires the



(a) Target Signal - Siren (b) Background Noise

Fig. 8. Gammatonegrams of the re-synthesised signals after segregation. Original gammatonegram of the sound mixture is provided in Figure 3b.

knowledge of the premixing sound sources (target signal and background noise separately), which is not available when dealing with realistic scenarios, such as the ones analysed in this paper. Alternative methods are based on listening tests and on visual examination of the re-synthesised sounds (both noise and target signal). An example of the latter is provided in Figure 8. The original gammatonegram of the sound mixture is given in Figure 3b. We can observe that the majority of the noise present in the sound mixture has been removed.

If more than one microphone is available, as in our system, the waveform of the target signal can be reconstructed for both channels and the relative sound source localised, using one of the methods described in the literature (for a review, see [40]).

V. CONCLUSIONS

In this paper we present a framework for alerting sound event detection and recognition in a driving scenario. The goal is to provide smart vehicles with the capability to interpret the urban soundscape and to react accordingly to specific acoustic events, which cannot be perceived by different sensing modalities. As we are dealing with realistic scenarios, characterised by non-stationary unstructured noise, we introduce a new noise-removal technique to obtain noise-free representation of the signals to use for classification. This representation relies on the idea of extracting the target signal, by processing the visual appearance of the mixture spectrogram. When compared to traditional feature representations such as Mel-frequency cepstrum coefficients, our framework shows an improvement up to 31% in the classification rate. Moreover, differently from the other methods, our approach provides a high classification accuracy that

is balanced across all classes considered. Future investigations could focus on the development of place-dependent soundscape models leading also to the detection of other anomalous sounds. Moreover, the auditory information could be combined with the visual appearance of the sound source for multi-modal detection and recognition.

ACKNOWLEDGMENT

The authors gratefully acknowledge support from the EU project EUROPA2 (FP7-610603).

REFERENCES

[1] M. Hersh and M. A. Johnson, *Assistive technology for visually impaired and blind people*. Springer Science & Business Media, 2010.

[2] 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.

[3] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," in *Advances in neural information processing systems*, 2005, pp. 547–554.

[4] L. Marchegiani and X. Fafoutis, "A behavioral study on the effects of rock music on auditory attention," in *International Workshop on Human Behavior Understanding*. Springer, 2013, pp. 15–26.

[5] X. Zhuang, X. Zhou, T. S. Huang, and M. Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 17–20.

[6] D. Chakrabarty and M. Elhilali, "Abnormal sound event detection using temporal trajectories mixtures," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 216–220.

[7] C.-F. Chan and W. Eric, "An abnormal sound detection and classification system for surveillance applications," in *Signal Processing Conference, 2010 18th European*. IEEE, 2010, pp. 1851–1855.

[8] L. Rabiner and B. Juang, "An introduction to hidden markov models," *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[9] 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 *Signal Processing Conference, 2008 16th European*. IEEE, 2008, pp. 1–5.

[10] J. Schröder, S. Goetze, V. Grutzmacher, and J. Anemüller, "Automatic acoustic siren detection in traffic noise by part-based models." in *ICASSP*, 2013, pp. 493–497.

[11] 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 ICROS-SICE International Joint Conference 2009*. IEEE Xplore, 2009.

[12] B. Widrow and S. D. Stearns, "Adaptive signal processing," *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1985, 491 p.*, vol. 1, 1985.

[13] L. Marchegiani, F. Pirri, and M. Pizzoli, "Multimodal speaker recognition in a conversation scenario," in *International Conference on Computer Vision Systems*. Springer, 2009, pp. 11–20.

[14] M. Janvier, X. Alameda-Pineda, L. Girinz, and R. Horaud, "Sound-event recognition with a companion humanoid," in *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*. IEEE, 2012, pp. 104–111.

[15] H. D. Tran and H. Li, "Sound event recognition with probabilistic distance svms," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 6, pp. 1556–1568, 2011.

[16] K. Łopatka, P. Zwan, and A. Czyżewski, "Dangerous sound event recognition using support vector machine classifiers," in *Advances in Multimedia and Network Information System Technologies*. Springer, 2010, pp. 49–57.

[17] H. Deshpande, R. Singh, and U. Nam, "Classification of music signals in the visual domain," in *Proceedings of the COST-G6 Conference on Digital Audio Effects*. sn, 2001, pp. 1–4.

[18] T. Dutta, "Text dependent speaker identification based on spectrograms," *Proceedings of Image and vision computing*, pp. 238–243, 2007.

[19] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[20] 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, 2013.

[21] C. K. Williams and C. E. Rasmussen, "Gaussian processes for machine learning," *the MIT Press*, vol. 2, no. 3, p. 4, 2006.

[22] K.-S. Fu and J. Mui, "A survey on image segmentation," *Pattern recognition*, vol. 13, no. 1, pp. 3–16, 1981.

[23] 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.

[24] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[25] L. Marchegiani, S. G. Karadogan, T. Andersen, J. Larsen, and L. K. Hansen, "The role of top-down attention in the cocktail party: Revisiting cherry's experiment after sixty years," in *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, vol. 1. IEEE, 2011, pp. 183–188.

[26] G. Shakhnarovich, T. Darell, and P. Indyk, "Nearest-neighbors methods in learning and vision," 2006.

[27] M. Weintraub, "A theory and computational model_ of auditory monaural sound separation," Ph.D. dissertation, Stanford University, 1985.

[28] T. W. Stokes, "Improving the perceptual quality of single-channel blind audio source separation." Ph.D. dissertation, University of Surrey, 2015.

[29] J. C. Murray, H. Erwin, and S. Wermter, "Robotics sound-source localization and tracking using interaural time difference and cross-correlation," in *AI Workshop on NeuroBotics*, 2004.

[30] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22st ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014.

[31] 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, 2010, pp. 3809–3812.

[32] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988.

[33] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[34] A. F. Pour, M. Asgari, and M. R. Hasanabadi, "Gammatonegram based speaker identification," in *Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on*. IEEE, 2014, pp. 52–55.

[35] V. Hautamäki, I. Kärkkäinen, and P. Fränti, "Outlier detection using k-nearest neighbour graph." in *ICPR (3)*, 2004, pp. 430–433.

[36] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, no. 2. ACM, 2000, pp. 93–104.

[37] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *ACM SIGMOD Record*, vol. 29, no. 2. ACM, 2000, pp. 427–438.

[38] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2002, pp. 15–27.

[39] R. Smith, A. Bivens, M. Embrechts, C. Palagiri, and B. Szymanski, "Clustering approaches for anomaly based intrusion detection," *Proceedings of intelligent engineering systems through artificial neural networks*, pp. 579–584, 2002.

[40] S. Argentieri, P. Danès, 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.