A Small-Footprint Deep-Learning Solution for Real-Time In-Car Emergency Vehicle Detection

Francesco Bossio*, Paolo Bestagini*, Luca Menescardi†, Federico Maver†,
Daniele Foscarin†, Michele Buccoli† and Simone Pecorino†
* Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano - Milan, Italy
† BdSound S.r.l.- Milan, Italy

***Abstract*—Modern vehicles emphasize cabin quietness and high-quality audio, which can mask critical auditory cues like emergency sirens. This issue is especially challenging for people with hearing loss. While some manufacturers have proposed automatic siren detection systems, existing solutions typically rely on external microphones, introducing hardware constraints. In this work, we tackle the problem of in-cabin siren detection, where low signal-to-noise ratios and spectral overlap with speech present significant challenges. We propose a novel, small-footprint detection system designed for real-time, frame-by-frame processing. Our system is trained on a synthetically generated dataset, enabling it to recognize diverse global siren sounds and generalize beyond region-specific recordings. Performance evaluations demonstrate its suitability for vehicle integration, offering a critical safety enhancement in increasingly quiet automotive environments.**

***Index Terms*—Siren detection, edge AI, acoustic event detection, in-cabin sensing, deep learning**

I. Introduction

The automotive industry is evolving along two parallel trends: enhancing cabin comfort through Active Road Noise Cancellation (ARNC) and improved Noise, Vibration, and Harshness (NVH) characteristics [1], while simultaneously refining the in-car audio experience with high-fidelity sound systems and personal listening zones [2]. While these advancements elevate passenger comfort, they also introduce a critical safety concern: masking essential auditory cues such as sirens from emergency vehicles. This issue is exacerbated in people with hearing loss, as evidenced by studies demonstrating a heightened risk of dangerous situations among elderly individuals with hearing impairment [3].

To address this, existing Emergency Vehicle Detection (EVD) systems typically rely on externally mounted microphones, which capture clearer audio signals by minimizing interference from in-cabin noise [4], [5]. However, this approach presents challenges for hardware integration, including weatherproofing, and secure mounting [6].

This work focuses on addressing in-cabin EVD, which presents two major acoustic challenges: (i) *Low Signal to Noise Ratio (SNR)* due to noise sources such as engine, tire and fan noise, and sound reflections; (ii) *Spectral overlap with speech and music* that share harmonic characteristics with siren sounds, making simple frequency-based separation, such as [7], ineffective and requiring more advanced analysis to distinguish between them. On the one side, large neural networks can achieve high accuracy by requiring significant computational resources. On the other side, computationally-light approaches often focus on detecting a specific type of siren, e.g., for a specific country [8]–[10].

In this work, we propose a small-footprint, low-resource approach capable of real-time general EVD. We train it on a synthetically generated dataset that includes a diverse range of siren sounds from various emergency vehicles worldwide (e.g., police cars, ambulances, fire trucks) and incorporates multiple international siren standards to ensure broad adaptability and avoid regional limitations. Synthetic data generation also allows for controlled and diverse training scenarios, addressing the scarcity of real-world siren recordings [11].

Alongside classification of the presence of an emergency vehicle, we add a layer for Voice Activity Detection (VAD), and we analyze its use to help disambiguation between speech and siren sources. We compare the performance of our approach against a baseline from the state of the art with traditional metrics from EVD. We evaluate the performance against different types of sirens. As ablation studies, we test the effectiveness of the VAD layer by removing it and testing the performance with a dataset with and without speech samples, and confirm the choice of a compact input audio representation.

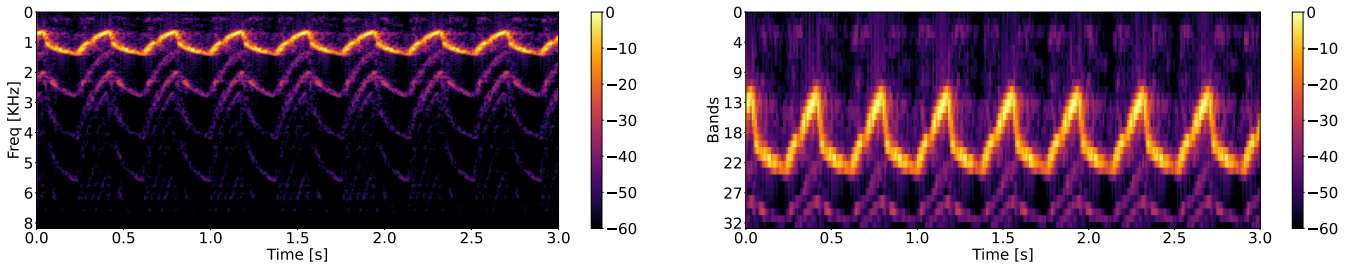
II. Problem Formulation

Modern vehicles are equipped with multiple microphones whose placement varies across different car models [12]. To ensure generality, in this study we focus on a scenario in which the signal $x[t]$ is acquired by a single microphone. The n-th frame of T samples of the acquired signal, capturing contributions from multiple acoustic sources, is denoted as

$$\mathbf{x}[n] = \mathbf{s}[n] + \mathbf{e}[n] + \mathbf{c}[n] + \mathbf{v}[n], \tag{1}$$

where $\mathbf{x}[n] = [x[nT], x[nT+1], ..., x[(n+1)T-1]]$; $\mathbf{s}[n]$ denotes one or more speech sources (e.g., driver, passengers, etc.); $\mathbf{e}[n]$ is the possibly occurring emergency vehicle siren; $\mathbf{c}[n]$ represents the audio playback from the car's loudspeakers (e.g., music, far-end phone calls, etc.); $\mathbf{v}[n]$ accounts for any other additional environmental noise (e.g., traffic, road noise, engine noise, etc.).

The signal frame $\mathbf{x}[n]$ includes any transformation introduced by the cabin's acoustic properties, such as room impulse



(a) Log-Energy spectrogram

(b) Input audio representation

Fig. 1: Comparison between the original spectrogram and the input of the neural network.

response for speech sources, and the Doppler effect for moving emergency vehicles. We assume the known playback signal $\mathbf{c}[n]$ is removed with an echo cancellation module and its contribution is negligible (i.e., $\mathbf{c}[n] \approx 0$).

The goal of real-time EVD is to detect the presence of an emergency vehicle siren at each acquired frame $\mathbf{x}[n]$, i.e., $|\mathbf{e}[n]| > 0$. To achieve this, we employ a deep learning technique to infer the probability of presence from the observed signal, i.e., $\hat{p}[n] = \mathcal{F}(\mathbf{x}[n]) \in [0, 1]$. A final binary classification decision is obtained by thresholding the estimated probability.

III. Types of sirens

The siren signal $\mathbf{e}[n]$ varies with the type of emergency vehicle and their country, as different countries have different legislation. In general, a siren sound is formalized as a periodic function Φ (often a square function [13]) of a fundamental frequency $f[t]$ as $e[t] = \Phi(f[t], t)$, where t is the discrete-time index.

The fundamental frequency function $f[t]$, which is time-varying, depends on the type of siren. The most common types of siren are named *wail*, *yelp*, and *two-tone*, each characterized by specific modulation behaviors [14]. While in wail and yelp sirens $f[t]$ spans the whole range of frequencies between f_{low} and f_{high}, in two-tone sirens $f[t]$ just alternates between f_{low} and f_{high}, whose values depend on country regulations.

In wail and yelp sirens, it takes T_{rise} and T_{fall} seconds to modulate between the two frequencies in the rising and the falling phase, respectively, with modulation patterns such as linear, quadratic or exponential variations. Wail sirens have T_{rise} and T_{falls} of several seconds, and yelp sirens have shorter times and usually higher frequencies [13], as seen in Fig. 1a.

In two-tone sirens, $f[t]$ is a square wave defining the cycle through which the two frequencies are sustained.

IV. Proposed approach

To correctly detect sirens in real-time audio sequences, we implement a neural network model starting with a compact audio representation to perform EVD at frame level.

Audio Representation. We compute the energy spectrogram of the signal, and filter it with a bank of 32 triangular filters equally spaced in the log-frequency domain between 0 and 3.5 kHz, focusing on the fundamental frequency $f[t]$ and making this solution suitable even for Narrow Band (NB) applications. Fig. 1 shows an example of a log-energy spectrogram of a yelp siren recording and its corresponding representation.

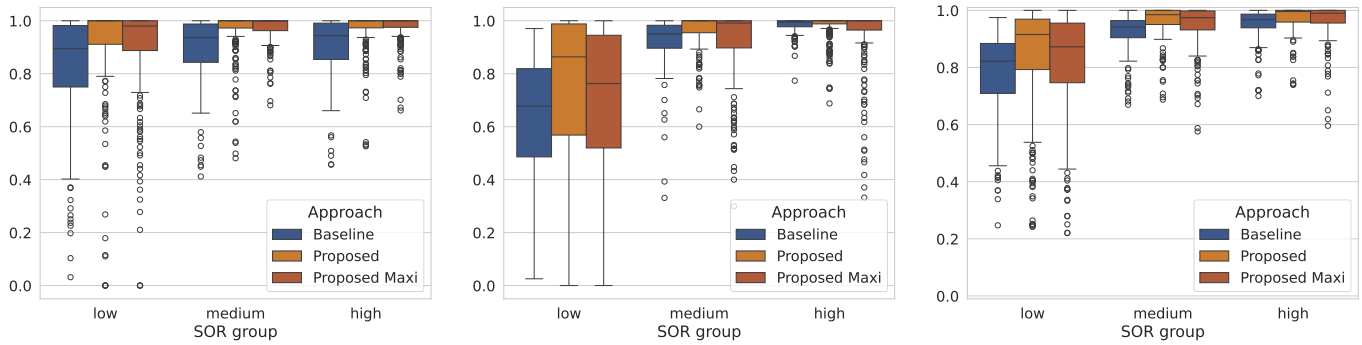
TABLE I: Network architecture of the proposed method

Layer	# pars	Output dim. (ch. first)	Channels
Conv 1 + MP	22	[2,16]	2
(Dep)Conv 2 + MP	60	[4,8]	4
(Dep)Conv 3 + MP	152	[8,4]	8
GRU	1008	[8]	8
FC_{siren}	9	[1]	
FC_{vad}	9	[1]	
Total	**1260**		

Network architecture. In Table I we summarize the network by reporting its layers and their number of parameters. In total, it employs 1260 parameters and requires only 0.25 Mega Multiply-Accumulate per second (MMACs) to work, making it a small model both in memory and computational requirements. We employ three depth-separable convolutional layers with 3×3 kernels followed by a batch normalization layer and a Rectified Linear Unit (ReLU) activation function and a max pooling layer over the frequency axis. These convolutional layers are trained in a causal fashion to make each output depending only on current and past frames, avoiding look-ahead. We then use a Gated Recurrent Unit (GRU) layer to capture long-term temporal dependencies from the extracted features, improving the model's ability to detect repeating siren patterns. Lastly, we employ two fully-connected classification layers, one for EVD and one for VAD. The two layers receive the same GRU's output in order to enhance the network's ability to disambiguate between sirens and speech. For this reason, the VAD layer is only used in training and it is not computed at inference time.

Data generation. Traditional EVD task does not require to disambiguate from speech and therefore common datasets for the task are not annotated with voice presence. Moreover, the variety of sirens is limited to the regulations of the countries where the recordings were acquired [15].

In order to train our model with robustness to speech and



(a) Precision (b) Recall (c) Accuracy

Fig. 2: Comparison on Precision, Recall, and Accuracy on the *real evaluation dataset* at different SORs.

effectiveness on different sirens, we build a dataset by mixing various audio samples to simulate the acoustic scenario, and we synthesize the sirens following the principles drawn in Section III as in [11].

We apply Doppler effect to generated sirens using the Pyroadacoustics library [16], mix them with noise samples at different gain and apply a band-pass filter to simulate the sound shielding effect of car windows. We then add speech samples and simulate different microphones frequency responses by means of biquad filters as in [17].

Training. The training of the neural network is performed with an *online* dataset, where a set of sequences is generated for each training step. This grants more variability to the training set, thus improving the generalization capability of the network. The training procedure is highly randomized, and each component is mixed with a random SNR $= |s|^2/|v|^2 \in$[0, 30] dB, a random Siren to Other Ratio (SOR) $|e|^2/|v+s|^2 \in$ [-15, 15]. Each mixed sample is then rescaled with a random gain between -15 and 0 dB, to train the network to detect sirens at varying microphone sensitivities.

The training loss is a weighted sum of binary cross-entropy losses for EVD and VAD, with weights of 10 and 1 empirically determined from the validation set, that has been generated with the same technique of the training set from unseen data.

V. Experimental Setup

The proposed network is trained with 10 s-sequences in batches of 8 sequences and with 128 steps per epoch, i.e., 2.84 hours of randomly generated audio sequences per epoch. For the audio representation we use 20 ms Vorbis windows with 10 ms stepsize ($T = 160$) as in [17], which allows the model to run EVD at 100 Hz.

The training set is generated using speech samples from the VCTK dataset [18], the noise samples from the DNS4 [19] dataset, sirens synthesized following the EN UNI 1789 for two-tone sirens and the SAE j1849, GSA K and CCR regulations for wail and yelp sirens [13].

Using different speech, noise and siren samples from the same datasets we also generate a *synthetic evaluation dataset* to maintain the control of the siren types.

We also generate a *real evaluation dataset* using different datasets to evaluate our proposed solution in out-of-domain conditions. We use speech samples from the EARS dataset [20], environmental noise samples from the UrbanSound8k dataset [21], road noise samples from the sireNNet dataset [22] and sirens from a collection of recordings covering emergency vehicles from various countries, including sirens whose type was not seen during training.

The synthetic and real evaluation datasets are composed of 1000 10-second audio sequences each.

VI. Evaluation

We evaluate our approach considering frame level classification, while skipping the excerpt during the 250 ms after siren presence change, in order to compare it with average human reaction time [23].

The approaches in the literature commonly perform EVD at excerpt level [4], [5]. In order to perform a fair comparison, we use as *Baseline* the approach from [5], adapted to work at frame-level by modifying the max pooling and flattening layers. This leads to reducing the size of the model from about 25 K parameters to about 4.5 K. We also compare our method with an extremely larger version of our approach, with about 4.7 M parameters, that we name *Proposed Maxi*. We avoid comparing our solution with larger self-supervised learning models for Sound Event Detection, as they do not comply with the computational requirements for EVDs in commercial applications.

In Fig. 2 we show the precision, recall and accuracy for three SOR ranges: low for SOR ≤ -5 dB, medium for $-5 <$ SOR < 5 dB and high for SOR ≥ 5 dB, against the real evaluation dataset. We can see that the proposed approach outperforms the baseline for every metric and for every SOR. We also observe that increasing the size of the proposed approach does not yield better results. This is likely due to the compact representation limiting the network's predictive power.

While at medium and high SOR the proposed approach achieves high results, at low SORs the model predictions lean toward higher precision and lower recall. This behavior can

TABLE II: Average metrics against the *real evaluation dataset*

	A	P	R	F1	EER
Baseline	87.01%	87.34%	81.41%	82.10%	13.07%
Proposed	**91.52%**	**94.61%**	**86.68%**	**87.79%**	**8.61%**
Proposed Maxi	89.69%	94.47%	82.77%	85.89%	9.15%

be tuned by choosing a lower threshold. It is worth highlighting that a negative SOR scenario is only expected with the emergency vehicle being extremely far from the vehicle, when a false negative may not be critical. With approaching emergency vehicles, the SOR is expected to increase at a higher level (as it is designed to be perceived by drivers [24]) and hence to be detected by the system.

In Fig. 3 we show the Receiver Operating Characteristic (ROC) curve and display the corresponding Equal Error Rate (EER). We notice again that the ROC curve of the proposed approach is always better than the baseline for every threshold, and even better than the large solution, achieving a final EER of 8.6% between false alarms and false reject. As EVD is a safety-focused task, we may want to reduce the False Reject Rate to 5%, which would require having a False Alarm rate of 20%. This is a general drawback of the solution.

In Table II we summarize the metrics against the whole evaluation dataset. We confirm that the Proposed approach is the best performing one, and that Recall is lower, leading to a lower F1 with respect to the accuracy.

Our approach is designed to be robust to different siren types and regulations. We verify this claim by analyzing the performance against the synthetic evaluation dataset over different types of sirens. The Proposed Maxi solution performs marginally better with wail and two-tone sirens, while the Proposed solution performs better with yelp sirens. We assume that this is caused by the larger GRU in the Proposed Maxi solution which may be more capable of learning long-duration patterns, such as those from the wail sirens and from the two-tone sirens. The yelp sirens, instead, are characterized by shorter temporal patterns, helping the more compact Proposed solution to perform better than the others. It is worth highlighting that while the EER for the two-tone sirens is significantly lower than that of the wail and yelp sirens, the other metrics are comparable. This suggests that the default threshold (0.5) is not optimal for the two-tone sirens and that with dedicated tuning, the performance may increase even more.

In general, the performance against the synthetic and the real evaluation datasets are comparable, confirming that the generation process creates realistic audio files.

VII. Ablation Studies

Firstly, we evaluate the effectiveness of the compact audio representation with respect to a network trained with a full log-energy spectrogram as shown in Fig. 1. From Table IV we see that using a full representation only guarantees a marginal improvement for Recall, F1 and EER scores, while achieving even lower Accuracy and Precision. We assume that the small

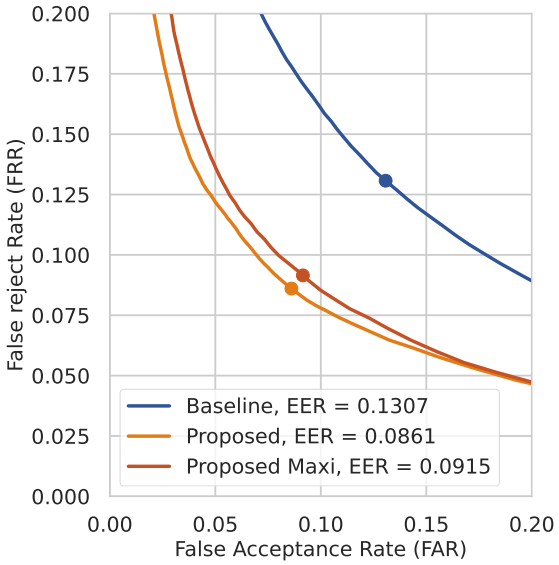


Fig. 3: ROC curves and corresponding EER values

TABLE III: Average metrics against the *synthetic evaluation dataset* at different siren types

Wail siren	**A**	**P**	**R**	**F1**	**EER**
Baseline	89.34%	82.45%	88.97%	83.12%	8.28%
Proposed	93.58%	88.18 %	85.51%	84.69%	7.69 %
Proposed Maxi	**95.47%**	**88.88%**	**93.86%**	**90.05%**	**5.06 %**
Yelp siren	**A**	**P**	**R**	**F1**	**EER**
Baseline	87.08%	93.85%	86.73%	88.70%	9.61%
Proposed	**90.40%**	**96.54%**	**89.21%**	**90.66%**	**7.69%**
Proposed Maxi	89.57%	96.41%	88.29%	88.98%	7.85%
Two-tone siren	**A**	**P**	**R**	**F1**	**EER**
Baseline	95.62%	92.21%	95.68%	92.91%	3.15%
Proposed	95.81%	94.94 %	92.16%	92.70%	3.36%
Proposed Maxi	**97.53%**	**95.27%**	**96.82%**	**95.58%**	**2.46%**

network does not hold enough predictive power to exploit the more accurate audio representation.

Secondly, we assess the impact of the VAD layer on improving robustness against speech presence by comparing the performance of the proposed approach trained with and without the VAD layer. This comparison is made using both the real evaluation dataset and an alternate version where speech is not added to the mixture. From Fig. 4, it is interesting to note that the VAD classification layer helps to improve performance even when tested against a dataset without speech, achieving 6.76% EER. When adding speech to the dataset, performance decreases for all the approaches under test, effectively demonstrating the issue of correctly disambiguating between speech and sirens. Nevertheless, it is clear that the proposed approach when trained with the VAD layer achieves better performance, with a 8.61% EER. It is worth noticing that the proposed approach trained without the VAD layer still achieves better performance than the baseline

TABLE IV: Comparison between different input audio representations

	A	P	R	F1	EER
32 Freq. Bands	**89.1**%	**91.9**%	83.0%	84.6%	11.41%
Log-Spec.	88.1%	89.6%	**86.2**%	**85.5**%	**11.15**%

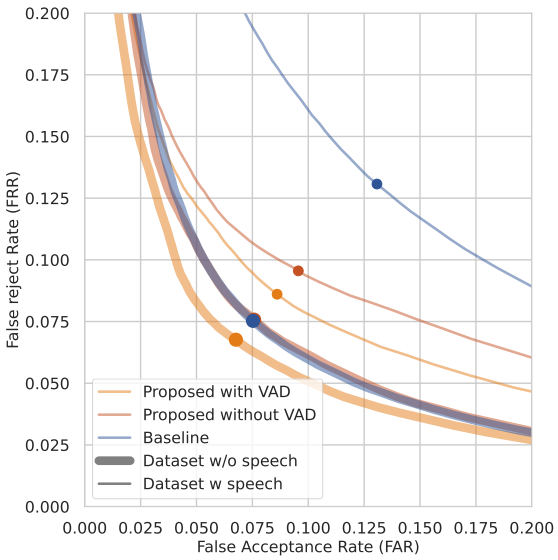


Fig. 4: ROC curves and corresponding EER values for VAD

approach, possibly due to its compact audio representation that focuses on the siren's fundamental frequencies and discards most of the harmonics that may lead to misclassification.

VIII. Conclusions

This paper presents a small-footprint, real-time siren detection system for the challenging in-cabin automotive scenario. We train our network on a synthetically generated dataset, enabling robust recognition of diverse global siren sounds, and include a VAD classification layer to provide robustness against speech presence. We compare our approach with a baseline solution and with a larger version of our proposed approach, achieving comparable or better results with respect to the baseline, especially in low SOR and against different types of siren. We also demonstrate the effectiveness of the VAD classification layer to help disambiguate between speech and siren and of the compact audio representations employed for our solution.

As future works, we intend to explore solutions to improve the performance of the approach in low SOR and reduce ROC curves and EER. We will also implement the proposed solution on a microcontroller-based device in order to verify its minimal computational complexity and performance in a real-world scenario.

References

[1] N. Zafeiropoulos, J. Zollner, and V. Kandade Rajan, "Active road noise cancellation for the improvement of sound quality in the vehicle," *ATZ worldwide*, vol. 120, no. 3, pp. 38–43, 2018.

[2] H. Oppenmann and S. Checa, "Sonic opportunities presented by personalized sound zones," in *Proceedings of the AES 4th International Conference on Automotive Audio*, 2022.

[3] L. Donmez and Z. Gokkoca, "Accident profile of older people in antalya city center, turkey," *Archives of gerontology and geriatrics*, vol. 37, no. 2, pp. 99–108, 2003.

[4] C.-L. Chin, C.-C. Lin, J.-W. Wang, W.-C. Chin, Y.-H. Chen, S.-W. Chang, P.-C. Huang, X. Zhu, Y.-L. Hsu, and S.-H. Liu, "A wearable assistant device for the hearing impaired to recognize emergency vehicle sirens with edge computing," *Sensors*, vol. 23, no. 17, 2023.

[5] M. Cantarini, A. Brocanelli, L. Gabrielli, and S. Squartini, "Acoustic features for deep learning-based models for emergency siren detection: an evaluation study," in *International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2021.

[6] J. Sieracki, M. Boehm, P. Patki, M. Caggiano, and M. Noll, "Seeing with sound: detection and localization of moving road participants with ai-based audio processing," in *In proc. of the AES 4th International Conference on Automotive Audio*, 2022.

[7] S. Damiano, T. Dietzen, and T. van Waterschoot, "Frequency tracking features for data-efficient deep siren identification," in *Proc. of the 10th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2024)"*, 2024.

[8] F. Beritelli, S. Casale, A. Russo, and S. Serrano, "An Automatic Emergency Signal Recognition System for the Hearing Impaired," in *2006 IEEE 12th Digital Signal Processing Workshop & 4th IEEE Signal Processing Education Workshop*, Sep. 2006.

[9] J.-J. Liaw, W.-S. Wang, H.-C. Chu, M.-S. Huang, and C.-P. Lu, "Recognition of the Ambulance Siren Sound in Taiwan by the Longest Common Subsequence," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2013.

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

[11] S. Damiano, B. Cramer, A. Guntoro, and T. van Waterschoot, "Synthetic data generation techniques for training deep acoustic siren identification networks," *Frontiers in Signal Processing*, vol. 4, 2024.

[12] F. Maver, D. Foscarin, D. Balsarri, L. Menescardi, M. Buccoli, S. Pecorino, and A. Grosso, "Audio speech source separation and enhancement in an automotive scenario using different microphone configurations," in *In proc. of the AES 5th International Conference on Automotive Audio*, 2024.

[13] R. Wagner, "Guide to test methods, performance requirements, and installation practices for electronic sirens used on law enforcement vehicles," 2000.

[14] V.-T. Tran and W.-H. Tsai, "Acoustic-based emergency vehicle detection using convolutional neural networks," *IEEE Access*, vol. PP, 04 2020.

[15] M. Asif, M. Usaid, M. Rashid, T. Rajab, S. Hussain, and S. Wasi, "Large-scale audio dataset for emergency vehicle sirens and road noises," *Scientific Data*, vol. 9, no. 1, Oct. 2022.

[16] S. Damiano and T. van Waterschoot, "Pyroadacoustics: a road acoustics simulator based on variable length delay lines," in *International Conference on Digital Audio Effects (DAFx)*, September 2022.

[17] J.-M. Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, 2018.

[18] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017.

[19] H. Dubey, V. Gopal, R. Cutler, S. Matusevych, S. Braun, E. S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "Icassp 2022 deep noise suppression challenge," in *ICASSP*, 2022.

[20] J. Richter, Y.-C. Wu, S. Krenn *et al.*, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Interspeech*, 2024.

[21] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *ACM International Conference on Multimedia (MM)*, 2014.

[22] A. Shah and A. Singh, "SireNNet-emergency vehicle siren classification dataset for urban applications," in *Mendeley Data*, 2023.

[23] C. Suied, P. Susini, and S. McAdams, "Evaluating warning sound urgency with reaction times," *Journal of Experimental Psychology. Applied*, Sep. 2008.

[24] K. Catchpole and D. McKeown, "A framework for the design of ambulance sirens," *Ergonomics*, vol. 50, no. 8, pp. 1287–1301, 2007.