Real-Time Acoustic Perception for Automotive Applications

Jun Yin*§, Stefano Damiano†§, Marian Verhelst*, Toon van Waterschoot†, Andre Guntoro‡⋆
*ESAT-MICAS KU Leuven, †ESAT-STADIUS KU Leuven, ‡Robert Bosch GmbH
⋆Email: andre.guntoro@de.bosch.com

***Abstract*—In recent years the automotive industry has been strongly promoting the development of smart cars, equipped with multi-modal sensors to gather information about the surroundings, in order to aid human drivers or make autonomous decisions. While the focus has mostly been on visual sensors, also acoustic events are crucial to detect situations that require a change in the driving behavior, such as a car honking, or the sirens of approaching emergency vehicles. In this paper, we summarize the results achieved so far in the Marie Skłodowska-Curie Actions (MSCA) Eruopean Industrial Doctorates (EID) project "Intelligent Ultra Low-Power Signal Processing for Automotive (I-SPOT)". On the algorithmic side, the I-SPOT Project aims to enable detecting, localizing and tracking environmental audio signals by jointly developing microphone array processing and deep learning techniques that specifically target automotive applications. Data generation software has been developed to cover the I-SPOT target scenarios and research challenges. This tool is currently being used to develop low-complexity deep learning techniques for emergency sound detection. On the hardware side, the goal impels workflows for hardware-algorithm co-design to ease the generation of architectures that are sufficiently flexible towards algorithmic evolutions without giving up on efficiency, as well as enable rapid feedback of hardware implications of algorithmic decision. This is pursued though a hierarchical workflow that breaks the hardware-algorithm design space into reasonable subsets, which has been tested for operator-level optimizations on state-of-the-art robust sound source localization for edge devices. Further, several open challenges towards an end-to-end system are clarified for the next stage of I-SPOT.**

***Index Terms*—Acoustic Perception, Sound Event Detection, Sound Source Localization, Embedded AI, Hardware-Algorithm Co-Design, Autonomous Driving, Microphone Array Processing**

I. The I-SPOT Project

The automotive industry is currently going through radical innovations that require rethinking the concept of mobility and the role of new technologies in the development of the automobile of the future. A key role in this evolving context is played by smart vehicles, aiming on one side to provide assistance to human drivers by enhancing their environmental awareness or supplying useful information in real-time, and on the other hand to achieve some level of autonomy in taking decisions and (partially) controlling the driving. At the time of writing, the first vehicles reaching level 3 of driving automation [1] begin to appear on the market, while manufacturers strive to achieve higher autonomy in future years, fueling the development of reliable technical solutions to ensure the safety of automated driving systems. Autonomous cars extensively rely on perceptual abilities, i.e. on the capacity to extract in real time information that is useful for driving safely, from data collected via different and multi-modal sensors mounted on the car. Most of the research on how to enhance the situational awareness of vehicles has been focusing on visual scene analysis, exploiting computer vision techniques and both long- and short-range imaging devices such as LIDARs, radars and cameras [2]. Human drivers, however, also strongly rely on *acoustic* cues as a way to perceive and understand what is happening in their surroundings: in certain scenarios, acoustic events are not accompanied by corresponding visual cues (e.g. a car honking) while in others they can enable the detection of objects that are invisible due to occlusions (e.g. an emergency vehicle behind a corner with an active siren). Acoustic perception can therefore complement the knowledge obtained from other sensory devices both in active (*drive*) and passive (*park*) mode, where sensing abilities can be exploited to continuously monitor the car and the surrounding environment for possible hazards (*always-on* operation). Among multiple potential use cases [3], highlighted in Fig. 1, such technology could help in: (i) detecting dangerous situations; (ii) identifying anomalies in car components; (iii) monitoring the acoustic scene for critical events.

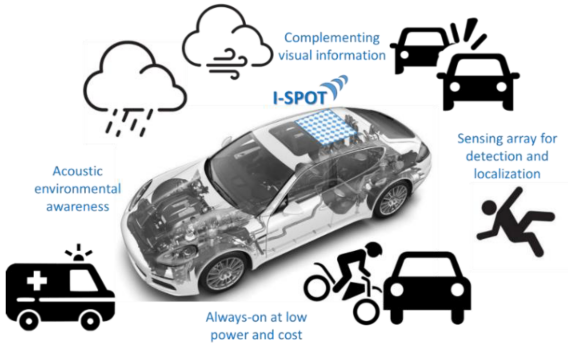


Fig. 1. I-SPOT targets the enhancement of smart cars with acoustic environmental awareness, through the addition of a smart sensing array capable of acoustic localization and detection, supported by custom processing hardware for efficient always-on operation.

To compensate for the current lack of acoustic awareness in cars and enhance the perception ability of autonomous vehicles, the EU MSCA project I-SPOT [4] targets the exploitation of acoustic sensing in the automotive domain. The project started in Nov. 2020 and relies on the collaboration between KU Leuven and Robert Bosch GmbH. Two early-stage researchers are involved in conducting research at the two partner institutions, working respectively on the algorithmic challenges and

§ These authors contribute equally to the paper.

on the hardware design, with the ultimate goal of merging the two components within this project. The project is now in its intermediate stage, and will run until Oct. 2024.

In this paper we describe the I-SPOT Project by introducing its motivation and goals, the results achieved during the first project stage and the problems still to be addressed. The rest of the paper is organized as follows: in Sec. II we introduce the target goals of I-SPOT, both on the algorithmic side and on hardware design aspects. Sec. III discusses the related research constituting the state of the art at the start of the project. Next, in Sec. IV we present what has been achieved so far and in Sec. V the work that still needs to be done during the second stage of the project. We finally draw conclusions in Sec. VI.

II. Project Goals

The ultimate goal of the I-SPOT Project is the enhancement of perceptual abilities of smart cars via acoustic sensing and signal processing. This broad task incorporates many sub-problems that can be grouped into algorithmic development and hardware design, tackled by the two researchers in an interleaved manner, where hardware assessment becomes an influential part of algorithm design and vice versa.

On an algorithmic design level, the project targets the development of low-footprint signal processing techniques for automotive acoustic signal identification, characterization and localization. This task corresponds to an application of the well-known Sound Event Localization, Detection (and Tracking), or SELD(t) problem in the automotive domain [5]. This use case brings several application-specific challenges to be addressed in the project:

1) The automotive acoustic scene is characterized by strong and dynamic background noise generated by multiple overlapping sound sources (e.g. other vehicles, people, environmental sounds such as rain or wind, etc.);
2) Due to the foreseen safety-critical use case, a high accuracy is required in both the sound event detection and localization;
3) The algorithms should be robust to the high variability of the sound events targeted in automotive applications (e.g. siren sounds are usually different in each country or region);
4) The target deployment on embedded devices requires the algorithms to be computationally and power-efficient: when operating in drive mode, a low-delay output is demanded, while an ultra-low power consumption is essential for the use in park mode;
5) To enable such efficient execution, algorithms cannot be developed in isolation from the embedded hardware design. Specifically, a hardware-algorithm co-design flow is required, which enhances the collaboration between the project partners.

Although the SELD(t) problem has been discussed at length in the literature and solutions have been proposed for both indoor (e.g. speaker identification [6], anomalous sound detection [7]) and outdoor applications (e.g. acoustic scene classification [8], traffic monitoring [7]), existing solutions do not cover these specific challenges of the automotive domain. Accommodating them constitutes one of the core innovative aspects of the I-SPOT Project.

Towards the hardware implementation and deployment, these algorithm targets also pose significant challenges. Firstly, hardware efficiency is critical in executing aforementioned computationally intensive algorithms in resource-restricted edge devices. Currently, the thriving field of AI deployment provides a wide range of hardware solutions for neural network execution [9]. However, the targeted algorithms require a hybrid approach with both neural networks in combination with conventional signal pre-/post-processing. Hardware platforms efficiently supporting such heterogeneous, diverse workloads are scarce and needed for power-efficient, low-latency end-to-end acoustic solutions [10]. Moreover, the automotive requirements asks for a reliable and well-packaged hardware system. Hence, on the hardware level, the I-SPOT system needs to feature:

1) Multi-kernel support for end-to-end heterogeneous algorithms;
2) Real-time low-latency operation to quickly response to each target event;
3) An optimized energy efficiency, to reduce the overall automobile power budget, especially in park mode.

Besides striving for efficiency under heterogeneous workloads, a second important goal of I-SPOT's hardware development, is to enable rapid design iterations and tolerate later algorithmic changes. In contrast to a traditional accelerator design flow in which the algorithm is considered stable when the accelerator development starts, the I-SPOT algorithm will evolve strongly during and even after hardware generation. We hence have to avoid on one hand that dedicated hardware architecture would reject new algorithmic features during the project, yet, also shy away from the development of too generic, flexible hardware architectures which lead to a lack of efficiency. This hence introduces several requirements for the I-SPOT hardware design workflow:

1) It is necessary to find a balance between flexibility and application-specific hardware architectures, to provide agile support of future algorithm upgrades.
2) The design procedure should enable rapid feedback for algorithm optimization via hardware assessment among multiple hardware design levels, especially the early design stage.
3) The workflow needs to feature hardware design reuse for fast integration and incremental module optimization.
4) Software-related optimization passes ought to form a scheduling tool that offers user-friendly programming of the resulting hardware.

The last system-level challenge is related to the assessment of the optimal microphone array topology and placement on the car body, in order to improve the quality of the received signal and boost the performance of the detection and localization algorithms. This task requires the definition of the desired number of sensors and their relative position, two parameters that can strongly influence the localization algorithms [11]. While the problem has been addressed in room scenarios, few

works concerning the automotive domain [12] exist. In such applications, only few places can effectively be used, due to the requirements of car manufacturers and the necessity to protect the sensors from the harsh environment, where the presence of strong vibrations, wind, temperature changes and atmospheric phenomena could cause damage and harm their correct operation.

To summarize, the I-SPOT Project embraces the design of an end-to-end system to provide autonomous cars with acoustic perception abilities. The foreseen challenges elaborated in this section link to a wide range of related research addressing the microphone array topology, the detection and localization of critical events, and efficient agile hardware architecture.

III. Related Work

Despite the wide literature on acoustic scene analysis, works targeting automotive applications are still limited. Some authors have tackled the detection and localization of emergency sound events, such as car horns and emergency sirens in an urban scenario [13]–[19]. These works can be split into the ones addressing only the detection problem [13], [14], [16], [17] and the ones targeting localization as well [15], [18], [19]. The proposed solutions to both problems are mostly based on end-to-end machine learning and deep learning methods, that have proven to provide an increased robustness to strong background noise and complex dynamic acoustic scenes as compared to traditional signal processing techniques [15]. The state-of-the-art approaches share a similar processing pipeline for the detection stage, eventually followed by a localization stage. The detection process can be summarized as follows. First, a feature-extraction step takes as input the audio signal recorded by one (for the detection-only problem) or multiple (if the localization is addressed too) microphones and builds a corresponding representation to be used as input to a neural network. Most methods use time-frequency feature representations, such as spectrograms [13], [14], [16], [19], gammatonegrams [14], [15], Mel-frequency cepstral coefficients (MFCCs) [13], [17], [19], or, less commonly, gammatone-frequency cesptral coefficients (GFCCs), constant-Q transform (CQT) and chromagrams [17]. Others take the raw waveform of the windowed audio signal as input feature [18]. Finally, some approaches adopt both time-frequency representations and the raw-waveform feature to train multi-path neural networks [13], [19]. After the feature extraction phase, the techniques for sound event detection and classification are mostly based on Convolutional Neural Networks (CNNs) [13], [14], [16], [17], [19], with the exception of [18], exploiting a fully-connected neural network, and [15], based on a U-net architecture [20] for the detection stage. The localization, instead, is tackled in [19] jointly with the detection, using an additional direction of arrival output added to the same network, while [15] includes a second CNN that takes as input the segmented gammatonegram features obtained during the detection stage. In [18], a traditional signal processing stage is cascaded to the detection network, in order to estimate both the sound's direction of arrival and distance.

Towards the realization of efficient models for the edge, the DCASE (Detection and Classification of Acoustic Scenes and Events) community [21] has recently obtained sub-100K-parameter models for low-complexity acoustic scene classification (DCASE2022 task-1), while the SELD problem (DCASE2022 task-3) still relies on over 10M neural network weights. The I-SPOT Project embodies such considerations in the algorithmic design, via the introduction of a co-design workflows to promote the joint optimization of model accuracy and complexity. This new research trend is still weakly explored in the audio signal processing research community [22], [23] as compared to the well-established research on network quantization and pruning for image processing using deep learning techniques [24]. This requires strengthening the interconnection between the heterogeneous algorithm development exploiting deep learning in combination with traditional signal processing on one hand, and versatile reconfigurable hardware architectures optimized for the automotive setting on the other hand.

The most straightforward approach towards hardware versatility across a wide range of algorithms is to make the hardware fully programmable. General-purpose devices such as CPUs and GPUs satisfy the demand of computational performance and runtime programmability for end-to-end hybrid algorithms, yet lack energy efficiency. More specialized devices like FPGAs could be controlled and configured at finer-granularity, yet still suffer from low power efficiency, and require considerable programming time. To overcome these bottlenecks, the modern coarse-grained re-configurable arrays (CGRA) [25] promise to offer a better balance between processor flexibility and energy efficiency. Several open-sourced CGRA platforms [26]–[29] have been proposed to accelerate heterogeneous workloads within certain application domains with more flexibility through run-time conditional controls. CGRA-based scheduling is explored at both fine and coarse hardware granularity, targeting to achieve either optimal utilization [26] through low level reconfigurability, or ultra-low power operation [30], [31] through specialized heterogeneous fabric design. Yet, the mapping algorithms for CGRAs remain challenging, and still fail to smoothly compile applications to modern CGRAs, especially when complex CGRA fabrics scale up.

IV. Achievements

During the first stage of the project, preliminary objectives have already been met, laying the foundations for the research tasks in the second phase towards the target goals presented in Sec. II.

A. Algorithmic Achievements

The first problem that has been addressed during the early phase of the project is the generation of data to design and assess algorithms for sound source detection and localization. Although some datasets containing urban sounds are available in the literature [32]–[34], they can hardly be used to address the I-SPOT challenges. First, these datasets target only the *detection* problem, thus usually providing single channel recordings with temporal labels. To perform sound source localization, instead, spatial information about the position and movement of sound sources in the acoustic scene is needed, and multi-microphone array data are required. Second, a tool enabling to flexibly generate data while changing the array

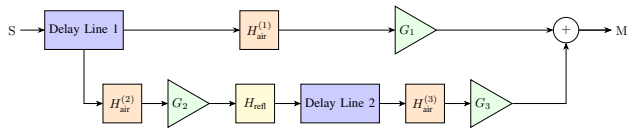


Fig. 2. *Pyroadacoustics* block scheme. Acoustic propagation is modeled using variable-length delay line elements; the sound attenuation caused by spherical propagation is implemented via three gain elements G_1, G_2, G_3; the asphalt reflection properties are modeled using an FIR filter H_{refl}; the air absorption properties are modeled via three FIR filters H_{air} [35].

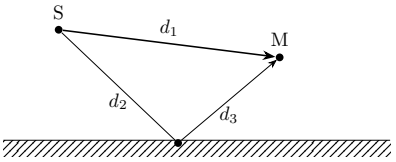


Fig. 3. Geometry of the simulated multi-path propagation in *pyroadacoustics*: the microphone M receives the direct signal emitted by the source S, propagating via d_1, and the reflected sound, via d_2 and d_3 [35].

configuration is of utmost importance to analyze the impact of different microphone array architectures on the detection and localization performance. This research topic is often missing in the literature, particularly in the automotive use case, probably also owing to the unavailability of sufficient data to support a systematic assessment.

These motivations led to the design of *pyroadacoustics*, a simulator of sound propagation specifically targeting road scenarios [35], that has been developed during the first stage of the project and released as an open-source Python library [36]. The simulator has a flexible design that enables the user to adapt the acoustic scene parameters to different use cases. Its architecture, depicted in Fig. 2, allows to simulate the sound produced by a single, omnidirectional sound source moving on an arbitrary trajectory with an arbitrary speed, and emitting a user-defined arbitrary audio signal. The sound is received by an array of an arbitrary number of omnidirectional, static microphones, placed in user-defined positions in the space, and includes both the direct component, and the reflection originating from the road surface (see Fig. 3). It should be noted that the support for user-defined trajectories ensures that the presence of both moving sources and microphones can be emulated by generating complex trajectories that account for variations in the relative speed between the source and the receivers (i.e. using spline or Bezier curves). To enhance the physical accuracy of the simulation, an accurate model of the reflection properties of the asphalt surface is included and implemented via FIR filtering, and can be adjusted by the user. Similarly, the effect of the air absorption is accounted for in the simulation via FIR filters. Finally, the use of variable-length delay lines to implement acoustic propagation ensures that the Doppler effect is correctly simulated [37]. This tool is, to the best of the authors' knowledge, the only one publicly available providing the possibility to generate multichannel audio data in a road scenario, and can be exploited to obtain a dataset for both the detection and localization of sound sources, as well as to conduct studies on microphone array geometries for automotive applications.

In order to kick-start the design of the envisioned algorithms during the second stage of the project, *pyroadacoustics* has then been used to generate a dataset for emergency vehicle detection. In order to include relevant sound events, audio samples containing clean sounds of different types of sirens (namely, hi-low, wail and yelp [15]) and car horns, all recorded with a close microphone in order to avoid any interference from other sources active on the road, have been collected from the online repository *www.freesound.com*, together with 2.5 hours of urban ambience noise and traffic noise. Exploiting these audio clips, a total of 15000 single-channel data samples have been generated using *pyroadacoustics*: each clip contains the sound emitted by a source of interest (i.e. either a siren or a car honking) moving on a random trajectory with an arbitrary speed, summed with a background noise extracted randomly from the collected noise clips. The sound and noise signals are combined with a random SNR in the range $[-30, 0]$ dB. This dataset is meant to be used for the training and assessment of sound event detection algorithms targeting automotive applications, while the same simulation tool will be exploited for the generation of data to target localization as well.

B. Hardware Flow Achievements

To pipe-clean the hardware-algorithm development workflow, early-stage hardware-based bottleneck assessments and optimizations have been conducted based on the Cross3D [38] sound source localization (SSL) framework, a state-of-the-art baseline. Cross3D combines conventional signal pre-processing, with a deep learning back-end to achieve efficient and accurate SSL. It specifically leverages pre-processing to extract the steered response power feature with phase transform filter (SRP-PHAT) and obtain robust SSL accuracy over unforeseen test data. In the back-end, Cross3D, replaces the typical hardware-unfriendly beam-forming computation for SRP-PHAT-based localization by a highly parallelizable CNN workload. In I-SPOT, we assessed the hardware benefits stemming from this optimization and further finetuned the Cross3D model towards more efficient edge deployment. To enable this, a hardware-oriented optimization workflow has been prototyped, as illustrated in Fig. 4, covering the following key features: (i) The bottleneck analysis on the baseline algorithms across the design parameter space definition; (ii) The algorithmic fine-tuning of design parameters based on resources including pre–trained model weights and empirical DNN training parameters; (iii) The multi-level hardware cost model evaluation that jointly considers different design stages, such as roofline analysis [39], PyTorch profiler, and TVM runtime performance [40]); (iv) The hardware-algorithm trade-off judgment from algorithm output and hardware overhead; (v) The global configuration update to narrow the design parameter space and handle the co-design porting exceptions.

This script-based workflow has proven its capability in squeezing Cross3D project for edge-device low-latency execution without giving in accuracy (8.59 ms/frame end-to-end on RasPi-4B, 7.26x faster than the baseline). The hardware-based analysis and low-complexity SRP literature [41] inspires a mathematically equivalent SRP-PHAT algorithm with $\sim$10x latency boost and $\sim$50% coefficients reduce. The algorithm-

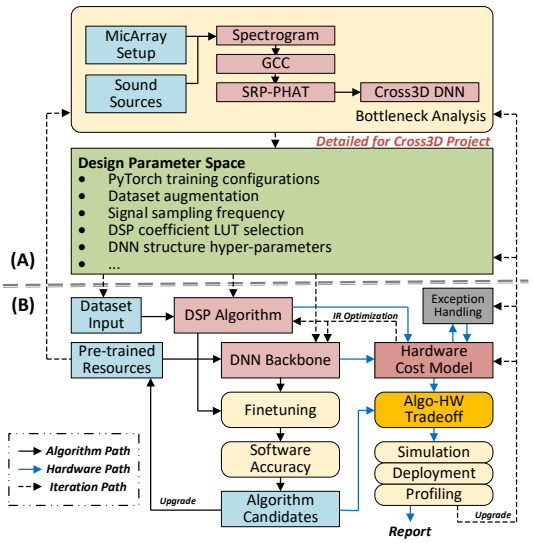


Fig. 4. Overview of the hardware-algorithm co-design workflow prototype for I-SPOT hybrid algorithms, with (A) algorithm-specific analysis and (B) general hardware-algorithm toolchain. The neural network part is powered by standard DNN frameworks. The hardware profiling branch relies on IR porting from the original algorithm descriptions to unified lower operator expressions (currently with the TVM IR library [40]).

hardware co-optimization helps to discover better training scripts and finetune the baseline model to edge-device versions which are ∼86% smaller while ∼47% faster.

V. Open Challenges

Based on the achieved milestones in Section. IV, we are eager to tackle the remaining bottlenecks during the second stage of the project. On the algorithmic side, the design of sound event detection and localization algorithms targeting emergency sounds in a road scenario (i.e. emergency sirens and car horns) is currently ongoing and will be further investigated in the next few months. This task, supported by the dataset described in Sec. IV, will again combine both deep learning and traditional signal processing techniques, towards a hybrid approach to solve the SELD(t) problem. This hybrid approach aims at increased hardware efficiency and enhanced explainability of the results, a crucial aspect for a safety-critical use case as autonomous driving. Hybrid approaches, combining spectral and spatial features for signal analysis, have proven effective for the localization of sound events produced by moving sources, as discussed in Sec. II, but they are still rarely adopted in the automotive domain [18].

On the hardware side, two major challenges are to be tackled in the next project stage. First, the script-based co-design workflow requires extra functionality to enable higher automation: During algorithmic development, hybrid algorithms typically contain multi-level scientific computation APIs , which differ in performance and hardware compatibility for porting and lowering. Hence, instead of handling these exceptions by manually selecting functions and rewriting the program towards better hardware mapping, an I-SPOT intermediate representation (IR) needs to be built. This will be pursued based on existing IR semantics, such as modifying the stateful data flow graph (SDFG) representation [42] targeting heterogeneous CGRA fabric backends. The toolchain prototype in Section. IV-B also leaves space for further IR customization that lowers high-level algorithm descriptions to actually generate custom, yet flexible hardware. Second, the I-SPOT hardware architecture will be defined in more detail. Currently, the co-design optimization iteration ends with hardware evaluation on embedded CPU processors. Along with the progress and survey from the software aspect, the first version of CGRA processing elements and hardware control blocks will be drafted for basic operators in the target algorithm. Moreover, with the flows currently in place, hardware design exploration and algorithm optimization will start to be more tightly interwoven.

Finally, on a system level, the design and assessment of the microphone array to be adopted for sensing purpose is yet to be tackled. This task also involves the assessment of the robustness of automotive SELD(t) methods to different microphone array geometries, justified by the need to potentially deploy these system on multiple classes of vehicles. The use of *pyroadacoustics* to generate data with different sensor array architectures makes such an assessment feasible.

VI. Conclusion

In this paper we introduced I-SPOT, an EU MSCA Project gathering two partner institutions, KU Leuven and Robert Bosch GmbH, that targets the enhancement of the acoustic perception of autonomous cars via audio signal processing. Within the project, the two aspects of algorithmic and hardware design co-exist in an interconnected development loop where hardware assessment requirements bring a foundational feedback into the algorithm design and tuning process, and vice versa. During the first stage of the project a road acoustics simulator has been designed, setting the ground for both the development of emergency sound event detection and localization algorithms, and the design and assessment of the sensor array needed to provide the vehicle with sensing capabilities. The first version of a hardware-algorithm co-design flow has been established via evaluating and optimizing a state-of-the-art hybrid SSL solution, along with the initial data collection of hardware bottlenecks and overhead levels towards the I-SPOT Project requirements. A hybrid approach has been chosen for the target audio signal processing algorithms, using a combination of traditional signal processing and deep learning techniques. This ensures an improved interpretability of the results compared to end-to-end deep learning methods, which is demanded by the safety-critical use case of autonomous driving. Remaining challenges are driving in our ongoing work and will be further addressed during the second stage of the project, together with (and with a continuous feedback from) the design of the edge-devices and accelerators on which the models will be deployed.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 956962 and from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only

the authors' views and the Union is not liable for any use that may be made of the contained information.

REFERENCES

[1] ISO/SAE PAS 22736, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," 2021.

[2] R. Hussain and S. Zeadally, "Autonomous Cars: Research Results, Issues, and Future Challenges," *IEEE Comm. Surv. Tutor.*, vol. 21, no. 2, pp. 1275–1313, 2019.

[3] L. Marchegiani and X. Fafoutis, "How Well Can Driverless Vehicles Hear? A Gentle Introduction to Auditory Perception for Autonomous and Smart Vehicles," *IEEE Intell. Transport. Syst. Mag.*, pp. 92–105, 2022.

[4] I-SPOT, "Intelligent Ultra Low-Power Signal Processing for Automotive," 2021. [Online], https://i-spot-project.eu/.

[5] S. Adavanne, A. Politis, and T. Virtanen, "Localization, Detection and Tracking of Multiple Moving Sound Sources with a Convolutional Recurrent Neural Network," in *Proc. Detect. Classif. Acoust. Scenes Events 2019 Workshop (DCASE2019)*, (New York), pp. 20–24, 2019.

[6] W. He, P. Motlicek, and J.-M. Odobez, "Deep Neural Networks for Multiple Speaker Detection and Localization," in *2018 IEEE Int. Conf. Rob. Autom. (ICRA)*, (Brisbane, QLD), pp. 74–79, 2018.

[7] Z. Mnasri, S. Rovetta, and F. Masulli, "Anomalous Sound Event Detection: a Survey of Machine Learning Based Methods and Applications," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5537–5586, 2022.

[8] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic Scene Classification: Classifying Environments from the Sounds They Produce," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 16–34, 2015.

[9] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Ai accelerator survey and trends," in *2021 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–9, IEEE, 2021.

[10] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[11] J. P. Dmochowski, J. Benesty, and S. Affes, "A Generalized Steered Response Power Method for Computationally Viable Source Localization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, 2007.

[12] O. Barak, N. Sallem, and M. Fischer, "Microphone Array Optimization for Autonomous-Vehicle Audio Localization Based on the Radon Transform," in *Proc. 5th Workshop Detect. Classif. Acoust. Scenes Events (DCASE 2020)*, (Tokyo), pp. 1–5, 2020.

[13] V.-T. Tran and W.-H. Tsai, "Acoustic-Based Emergency Vehicle Detection Using Convolutional Neural Networks," *IEEE Access*, vol. 8, pp. 75702–75713, 2020.

[14] M. Cantarini, A. Brocanelli, L. Gabrielli, and S. Squartini, "Acoustic Features for Deep Learning-Based Models for Emergency Siren Detection: An Evaluation Study," in *2021 12th Intern. Symp. Image Signal Process. Analysis (ISPA)*, (Zagreb, Croatia), pp. 47–53, 2021.

[15] L. Marchegiani and P. Newman, "Listening for Sirens: Locating and Classifying Acoustic Alarms in City Scenes," *IEEE Trans. Intell. Transport. Syst.*, vol. 23, no. 10, pp. 17087–17096, 2022.

[16] F. Walden, S. Dasgupta, M. Rahman, and M. Islam, "Improving the Environmental Perception of Autonomous Vehicles using Deep Learning-based Audio Classification," *arXiv:2209.04075 [cs.SD]*, Sept. 2022.

[17] J. Sharma, O.-C. Granmo, and M. Goodwin, "Emergency Detection with Environment Sound Using Deep Convolutional Neural Networks," in *Proc. 5th Int. Congr. Inf. Comm. Technol.*, (Singapore), pp. 144–154, 2021.

[18] Y. Furletov, V. Willert, and J. Adamy, "Auditory Scene Understanding for Autonomous Driving," in *2021 IEEE Intelligent Vehicles Symposium (IV)*, (Nagoya, Japan), pp. 697–702, July 2021.

[19] H. Sun, X. Liu, K. Xu, J. Miao, and Q. Luo, "Emergency Vehicles Audio Detection and Localization in Autonomous Driving," *arXiv:2109.14797 [cs.SD]*, Sept. 2021.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. Med. Image Computing Computer-Assisted Interv. (MICAAI)*, pp. 234–241, 2015.

[21] DCASE2022, "IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events," 2022. [Online], https://dcase.community/challenge2022/index.

[22] M. Mohaimenuzzaman, C. Bergmeir, and B. Meyer, "Pruning vs XNOR-Net: A Comprehensive Study of Deep Learning for Audio Classification on Edge-Devices," *IEEE Access*, vol. 10, pp. 6696–6707, 2022.

[23] G. Cerutti, R. Andri, L. Cavigelli, M. Magno, E. Farella, and L. Benini, "Sound Event Detection with Binary Neural Networks on Tightly Power-Constrained IoT Devices," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, pp. 19–24, Aug. 2020.

[24] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," *arXiv preprint arXiv:2103.13630*, 2021.

[25] A. Podobas, K. Sano, and S. Matsuoka, "A survey on coarse-grained reconfigurable architectures from a performance perspective," *IEEE Access*, vol. 8, pp. 146719–146743, 2020.

[26] M. Karunaratne, A. K. Mohite, T. Mitra, and L.-S. Peh, "Hycube: A cgra with reconfigurable single-cycle multi-hop interconnect," in *Proceedings of the 54th Annual Design Automation Conference 2017*, pp. 1–6, 2017.

[27] R. Bahr, C. Barrett, N. Bhagdikar, A. Carsello, R. Daly, C. Donovick, D. Durst, K. Fatahalian, K. Feng, P. Hanrahan, *et al.*, "Creating an agile hardware design flow," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, IEEE, 2020.

[28] J. Anderson, R. Beidas, V. Chacko, H. Hsiao, X. Ling, O. Ragheb, X. Wang, and T. Yu, "Cgra-me: An open-source framework for cgra architecture and cad research," in *2021 IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, pp. 156–162, IEEE, 2021.

[29] C. Tan, N. B. Agostini, J. Zhang, M. Minutoli, V. G. Castellana, C. Xie, T. Geng, A. Li, K. Barker, and A. Tumeo, "Opencgra: Democratizing coarse-grained reconfigurable arrays," in *2021 IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, pp. 149–155, IEEE, 2021.

[30] G. Gobieski, A. O. Atli, K. Mai, B. Lucia, and N. Beckmann, "Snafu: an ultra-low-power, energy-minimal cgra-generation framework and architecture," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, pp. 1027–1040, IEEE, 2021.

[31] M. Giordano, K. Prabhu, K. Koul, R. M. Radway, A. Gural, R. Doshi, Z. F. Khan, J. W. Kustin, T. Liu, G. B. Lopes, *et al.*, "Chimera: A 0.92 tops, 2.2 tops/w edge ai accelerator with 2 mbyte on-chip foundry resistive ram for efficient training and inference," in *2021 Symposium on VLSI Circuits*, pp. 1–2, IEEE, 2021.

[32] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *Proc. 22nd ACM Int. Conf. Multimedia*, (Orlando, USA), pp. 1041–1044, 2014.

[33] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, (New York, USA), pp. 1015–1018, 2015.

[34] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *2017 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 776–780, Mar. 2017.

[35] S. Damiano and T. van Waterschoot, "Pyroadacoustics: a Road Acoustics Simulator Based on Variable Length Delay Lines," in *Proc. 25th Int. Conf. Digital Audio Effects (DAFx20in22)*, (Vienna), pp. 216–223, 2022.

[36] S. Damiano, "pyroadacoustics - Python Road Acoustics Simulation," 2022. Available at: https://github.com/steDamiano/pyroadacoustics.

[37] J. O. Smith, *Physical Audio Signal Processing*. online book, 2010 edition, http://ccrma.stanford.edu/˜jos/pasp/, 2010.

[38] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust Sound Source Tracking Using SRP-PHAT and 3d Convolutional Neural Networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 300–311, 2021.

[39] S. Williams, A. Waterman, and D. Patterson, "Roofline: an insightful visual performance model for multicore architectures," *Communications of the ACM*, vol. 52, no. 4, pp. 65–76, 2009.

[40] T. Chen, T. Moreau, Z. Jiang, L. Zheng, E. Yan, H. Shen, M. Cowan, L. Wang, Y. Hu, L. Ceze, *et al.*, "{TVM}: An automated {End-to-End} optimizing compiler for deep learning," in *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pp. 578–594, 2018.

[41] T. Dietzen, E. De Sena, and T. Van Waterschoot, "Low-complexity steered response power mapping based on nyquist-shannon sampling," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 206–210, IEEE, 2021.

[42] T. Ben-Nun, J. de Fine Licht, A. N. Ziogas, T. Schneider, and T. Hoefler, "Stateful dataflow multigraphs: A data-centric model for performance portability on heterogeneous architectures," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–14, 2019.