Applied Acoustics 215 (2023) 109719

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

A safety-oriented framework for sound event detection in driving scenarios

Carlos Castorena *, Maximo Cobos, Jesus Lopez-Ballester, Francesc J. Ferri

Departament d'Informàtica, Universitat de València, 46100 Burjassot, Spain

ARTICLE INFO

Keywords:
Sound event detection
Deep learning
Driver safety
Smart vehicles
Sound event taxonomy

ABSTRACT

The safety of drivers has become increasingly important in today's rapidly evolving transportation landscape, especially with the rise of autonomous and smart vehicles. This paper proposes a safety-oriented framework for sound event detection in smart vehicles using deep learning models. The goal of this framework is to increase driver awareness, prevent accidents, and provide acoustic forensic analysis. To achieve this, a meaningful taxonomy of event classes in a driving scenario is introduced, taking into account the event classes that are known to be related to major driving distractors. Based on this taxonomy, a dataset has been created to train and evaluate a fully-convolutional sound event detection model that was inspired by the well-known YOLO vision model. Experimental results demonstrated that the proposed model offers competitive results, outperforming a state-of-the-art baseline using recurrent connections. This comprehensive framework for sound event detection in smart vehicles aligns with the recommended directions for future mobility scenarios and has the potential to significantly improve the safety and performance of smart vehicles.

1. Introduction

Road safety is of paramount importance in the contemporary world and constitutes a fundamental aspect of the sustainable development goals (SDGs) [1]. The SDGs establish ambitious objectives, such as Target 3.6, which endeavors to reduce by half the number of fatalities and injuries resulting from traffic accidents on a global scale. Moreover, Target 11.2 specifically aims to enhance measures pertaining to road safety. Although there has been a decline in the occurrence of car accidents, it is worth noting that they continue to rank as the eighth leading cause of death worldwide, comprising 2.5% of all fatalities [2]. Nonetheless, it is imperative to acknowledge that many of these accidents can be averted through the implementation of appropriate road safety practices.

In response to these challenges, researchers have increasingly explored the application of deep neural networks for accident prevention. For instance, the latest generation of autonomous vehicles leverages multiple sensors to make critical decisions, such as braking or steering wheel control [3]. Eye tracking technology has also been employed to detect instances when a driver is not paying adequate attention to the road [4]. Furthermore, generative networks have been utilized to mitigate the impact of fog in camera systems used for assisted driving [5]. Works such as [6] focus on accident prevention by detecting stress patterns in drivers and in the case of [7] it mixes different techniques such as object detection, tracking and distance measurement in order to provide information and prevent accidents. Additionally, there have been significant advancements in employing convolutional neural networks (CNNs) for image processing to detect distractions and activities that pose potential risks [8–10]. A recent review on learning-based traffic accident prediction models based on heterogeneous data sources can be found in [11].

Distractions can be categorized as external or internal stimuli that divert an individual's attention from the primary task or activity at hand. Within the realm of road safety, distractions encompass any factors that cause a driver to shift their focus away from driving, thereby potentially heightening the risk of accidents or compromising their capacity to respond to hazards in a timely manner. Distractions may manifest in diverse forms, including visual distractions (e.g. attractive scenery), auditory distractions (e.g. phone calls), cognitive distractions (e.g. emotional conversations), and physical distractions (e.g. adjusting in-car controls). Such distractions can substantially impede a driver's capability to perceive, process, and react to vital information during the operation of a vehicle.

Extensive literature highlights the mobile phone as a primary source of distraction during driving [12–15]. Whether it involves making phone calls or sending and receiving text messages, the use of mobile phones significantly elevates the risk of accidents, with studies indicat-

* Corresponding author.
E-mail address: carlos.castorena@uv.es (C. Castorena).

https://doi.org/10.1016/j.apacoust.2023.109719
Received 8 August 2023; Received in revised form 1 October 2023; Accepted 26 October 2023
Available online 9 November 2023

ing a 23-fold increase in the likelihood of mishaps [2,15]. Despite legal restrictions on mobile phone use while driving in approximately 150 countries [2], research conducted by Prat et al. [14] in Girona (Catalonia, Spain) reveals that approximately 40% of individuals in the region admit to using their mobile phones while driving, particularly among young people. Another crucial factor discussed in various studies is the presence of passengers [16–20]. Interacting with passengers, whether through conversations or attending to the needs of children or pets, is widely recognized by multiple authors as a significant source of distraction. Unlike mobile phone use, this aspect is not subjected to specific regulations and represents the most prevalent practice that can be deemed distracting while driving.

In addition to the aforementioned distractions, weather conditions and their interaction with the vehicle can create hazardous situations. While not considered a distraction in itself, weather significantly influences and escalates the risk of accidents. Cai et al. [21] highlight that the United States alone reports over one million weather-related accidents, with China documenting more than fifty thousand such incidents. While most studies on distraction detection primarily concentrate on visual cues, certain events are not easily discernible through visual means. For instance, situations like a ringing mobile phone or specific speech interactions among passengers may not be readily detectable through visual analysis alone. Consequently, the utilization of algorithms that leverage audio as input can aid in enhancing the detection of these distractions.

Sound event detection (SED) has emerged as a highly significant research domain with immense potential for integration into driving safety assistance systems. SED involves not only identifying the specific type of event occurring in an audio track but also accurately detecting the temporal occurrence of these events. The ability to recognize and classify sound events in real-time can greatly enhance driver safety by providing valuable information about the surrounding environment and potential hazards. Over the past decade, SED has gained substantial attention and experienced remarkable advancements. The ability to automatically recognize and classify a wide range of sound events in different contexts has implications for numerous domains, including public safety, smart cities, and human-computer interaction. Researchers have explored various scenarios, including domestic environments [22], urban environments [23], and more, each presenting unique challenges related to environmental acoustics and event characteristics. This research area has witnessed notable progress, particularly through the utilization of advanced deep learning techniques.

In this study, our primary focus is on investigating the detection of sound events for the purpose of identifying distractions, employing state-of-the-art SED models. The work makes several significant contributions. Firstly, we establish a taxonomy of common sounds that may be present in a driving scenario. The objective of this taxonomy is to define a comprehensive set of sound classes that are highly relevant to driving activities, with an emphasis on identifying potential distractors based on existing literature. Secondly, we introduce an open audio dataset specifically curated to train and evaluate SED systems within driving environments. The dataset is constructed based on the established taxonomy and encompasses a wide range of driving sounds extracted from multiple sources, simulating in-vehicle audio recordings. This comprehensive dataset enables the robust training and evaluation of SED models for driving safety applications. Lastly, we propose and assess the effectiveness of integrating the YOLO (You Only Look Once) architecture [24] into a SED-oriented assistance system. By leveraging the renowned YOLO framework, known for its efficiency and accuracy in object detection tasks, we aim to develop a fully-convolutional model seamlessly integrable into real-time driving safety systems, just like its vision counterpart. Through these contributions, our study aims to advance the field of SED in the context of driver safety, with the ultimate goal of enhancing awareness and reducing distractions for improved road safety.

Fig. 1. Driver exposed both to internal (e.g. ring tone) and external sounds.

The remaining sections of this paper are organized as follows. Section 2 explains and deepens the proposed taxonomy. Section 3 provides an overview of the data used in our study, including its description, acquisition process, labeling methodology, and processing techniques. Section 4 explores state-of-the-art methodologies for identifying sound events, with a focus on the application of YOLO as an effective strategy for SED. Section 5 introduces various metrics used to evaluate the performance of our proposed approach. Section 6 discusses the main findings of our study, analyzing the synthetic and real data separately to provide a comprehensive understanding of our approach's effectiveness and establishing a starting point for future research. Finally, Section 7 key outcomes of our research, highlights their implications for driver safety, and discusses potential directions for future studies in this field.

2. Taxonomy for driving sounds

In this section, we present a taxonomy for driving sounds, focusing on its significance in understanding the acoustic environment encountered during driving scenarios. The taxonomy aims to categorize and organize a wide range of sounds that may occur while driving, providing a comprehensive framework for analyzing and characterizing the auditory landscape within and around vehicles. By establishing this taxonomy, we aim to enhance our understanding of the acoustic aspects of driving, which has implications for various areas such as driver comfort, safety, and in-vehicle audio system design.

Driving sounds encompass a diverse set of auditory events that drivers are exposed to while operating a vehicle. These sounds can originate from both internal sources within the vehicle, such as vehicle components, devices, and human interactions, as well as external sources like traffic, road conditions, and environmental factors (see Fig. 1). Understanding the nature and impact of these sounds is crucial for creating pleasant and safe driving environments.

The taxonomy for driving sounds serves several important purposes. Firstly, it provides a structured framework for identifying, classifying, and categorizing different types of sounds encountered during driving. By organizing these sounds into meaningful classes and subclasses, the taxonomy enables a systematic analysis of their characteristics, sources, and potential effects on drivers. Secondly, the taxonomy facilitates the creation of labeled datasets for training and evaluating audio processing algorithms and systems related to driving. Accurate annotation of audio recordings with the corresponding classes and subclasses defined in the taxonomy enables the development of robust and reliable models capable of detecting and recognizing specific driving sounds. Furthermore, the taxonomy promotes effective communication and collaboration among researchers, practitioners, and stakeholders in the field

of automotive acoustics. With a common language and classification scheme provided by the taxonomy, discussions on driving sounds become more precise, allowing for the exchange of ideas, comparison of research findings, and benchmarking of audio processing techniques.

To ensure the validity and usefulness of the taxonomy for studying driving sounds, we have considered several factors during its development. Firstly, the taxonomy builds upon prior research and existing taxonomies to leverage existing knowledge and provide a comprehensive understanding of the acoustic environment in driving scenarios. This ensures a coherent and consistent framework that aligns with the existing body of literature [23,25,26]. Secondly, the taxonomy aims to capture the variety of sounds encountered during driving by incorporating multiple levels of classification. From the general location of the sound (inside or outside the vehicle) to the primary source of the sound (e.g., vehicle components, human interactions, environmental factors), and further to the specific event or sound produced, the taxonomy enables a hierarchical organization of driving sounds. Lastly, while the taxonomy covers a wide range of driving sounds, we specifically identify and highlight audio distractors within the taxonomy. These audio distractors are sounds that have the potential to divert driver attention and may pose safety risks. By identifying and analyzing these audio distractors within the taxonomy, we can gain insights into their characteristics, impacts, and potential countermeasures.

The subsequent subsections will provide a detailed description of the taxonomy, including the hierarchical structure, subclasses, and specific events encompassed within each level. This taxonomy will serve as a foundation for studying the acoustic aspects of driving and further exploration of audio distractors within the driving sounds framework.

2.1. Level hierarchy

The *Location Level* includes two main classes: Inside and Outside. Both classes lead to the next *Source Level*, representing the origin of subsequent classes. The Inside class represents sounds generated within the vehicle, including sources such as the vehicle itself, devices, people, and pets. These sounds directly affect the driver and occupants of the vehicle. On the other hand, the Outside class comprises sounds originating from external sources, including other vehicles, weather conditions, roads, and the surrounding environment. These sounds provide contextual information about the driving environment and may indirectly impact the driver's experience.

The *Specific Source Level* and the *Event Level* further categorize and provide a detailed understanding of the different sources of sounds encountered during driving scenarios. Building upon the *Source Level*, this hierarchical structure delves into the specific sources and events associated with the Inside and Outside classes.

2.2. Inside sounds

Under the "Car" source, we have "Car Elements", which includes ignition, engine, door, window, and wiper sounds, providing insights into various aspects of the vehicle's mechanical operations. "Alerts" represent warning beeps that may serve as auditory cues for the driver. Additionally, the "Audio System" involves a range of sounds such as music, ring tones, notifications, and speech, which can be emitted by the car's audio system or connected devices. The "Devices" source focuses on sounds related to electronic devices present in the vehicle. The "Audio System" subclass includes sounds like music, ring tones, notifications, and speech, which can originate from the vehicle's audio system or connected devices. The "Phone" subclass covers specific events such as ring tones, vibrating alerts, notifications, and speech that are directly associated with the phone itself. The "People" source captures sounds originating from human occupants in the vehicle. Interactions involve speech, representing conversations and communication among individuals. "Emotions" encompass a range of expressions like laughter, scream, sing, yell, and cry. "Physiological" events include nonverbal sounds like sneezing, coughing, and yawning, while "Activities" cover actions such as manipulating objects and drinking. "Pets", as a source, includes sounds produced by accompanying animals, such as dogs and cats, which can add to the acoustic environment within the vehicle. Lastly, the "Other" source focuses on the occurrence of knocking sounds, which may arise from various sources like hits and microphone noise.

2.3. Outside sounds

The Outside class encompasses a range of sources and events encountered outside the vehicle during driving scenarios. Under the "Vehicles" source, we have "Accidents", which include crash and drift events, providing insights into collisions and loss of vehicle control. "Passing" represents events associated with the presence of other vehicles, such as airplanes, helicopters, cars, trucks, trains, and motorcycles passing by. "Horns" capture different horn sounds, including those from boats, cars, trucks, and trains. "Sirens" encompass emergency vehicle sirens, including police, ambulance, and firefighters.

The "Atmospheric" source focuses on sounds related to various weather conditions. Weather events include thunder, rain, and hail, providing an understanding of the auditory environment during different weather patterns. The "Roads" source highlights sounds associated with different road surfaces. "Surfaces" encompass pavement, stone, and dirt, capturing the variations in the acoustic characteristics of the road itself. The "Environment" source provides a broader category that encompasses the overall acoustic environment outside the vehicle. It includes subcategories such as Urban Sound, Field, and Highway, which may connect to other existing taxonomies that specifically address sound events in those environments, like the Urban Sound taxonomy [23]. The Urban Sound taxonomy focuses on categorizing various sound events that are commonly encountered in urban settings, such as sounds from street traffic, construction sites, public transportation, and human activities. By linking to the Urban Sound taxonomy, the Driving Sounds taxonomy recognizes the importance of urban sound events and their potential impact on driver safety.

2.4. Auditory distractors

In the context of SED for driving safety, it is crucial to identify and prioritize auditory distractors that can significantly impact a driver's attention and pose risks on the road. The selection of these distractor classes has been informed by existing literature and previous work in the field [14,16,27–30]. While there are various factors that can divert a driver's attention, we specifically focus on auditory distractors in this discussion.

Within the taxonomy for Driving Sounds, the identified distractor classes can be found at both the *Event Level* and the *Specific Source Level*. The goal is to prioritize classes that are directly relevant to SED systems for driving safety, considering the specific requirements and objectives of such systems. In selecting these specific classes as auditory distractors, we prioritize their relevance to SED systems for driving safety rather than focusing on distinguishing between subclasses that may have minimal impact on driver attention. For example, while it may not be crucial to differentiate between specific types of pets or the exact source of horn sounds, it is essential to detect and recognize events such as specific interactions with a mobile phone device (which is widely recognized as a major source of distraction). The list of selected classes widely recognized as auditory distractors and of interest for a potential SED-based system is: `Ring Tone` (Phone/Audio System), `Vibrating`, `Notification` (Phone/Audio System) [12–15], `Speech` (Occupants/Audio System/Phone) [16,17,19,20], `Baby Cry` [31,27], `Physiological` [28,32], `Pets` [33,34,28], `Horns` [29] and `Sirens` [30]. Parentheses indicate that some event classes may be reproduced by two different specific sound sources inside the vehicle,

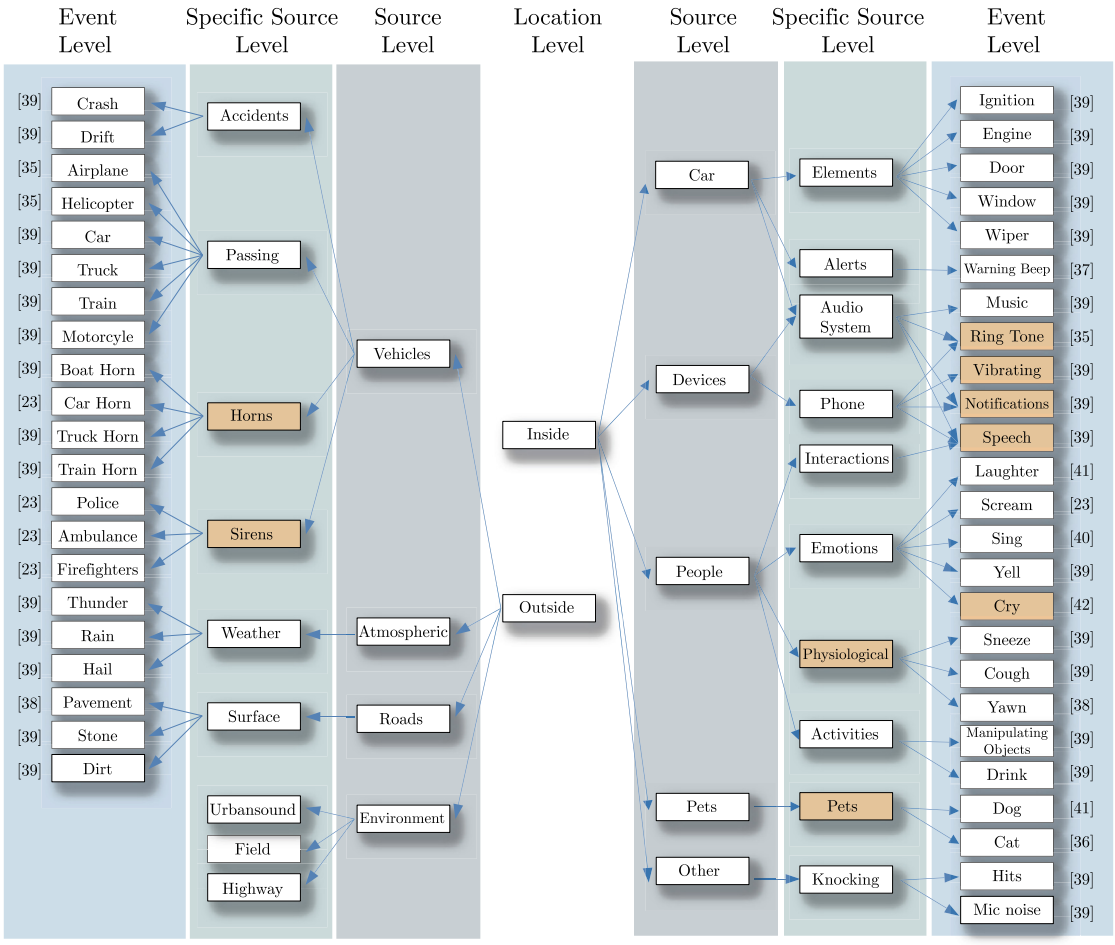


Fig. 2. Driving Sounds taxonomy with specific event sources as references. Selected auditory distractors are highlighted in orange.

which is an important aspect in the development of the dataset that will be described next. Therefore, in the forthcoming experimental sections of the present work, we focus on $C = 9$ distinct classes of interest, highlighted in orange in Fig. 2.

In considering the expansion of our dataset, we recognize the valuable input that medical experts in neurology, cognitive psychology, as well as professors or experts in the field of driving safety, can provide. By involving these specialists, we can gain a deeper understanding of which sounds may have a more pronounced impact on cognitive processes and driver behavior. Incorporating this broader perspective will contribute to a more holistic understanding of auditory distractors in driving scenarios.

3. Dataset

The dataset used in this study has been specifically designed to facilitate the training and evaluation of SED systems in the context of driving safety. The primary objective of this dataset is to enable the identification of auditory distractors, which have been identified as significant factors impacting driver attention and posing risks on the road. While the comprehensive taxonomy presented earlier encompasses all audible events usually encountered in driving scenarios, the focus of this dataset is on the auditory distractors highlighted in Fig. 2, as acknowledged in previous literature [14,16,27–30]. However, it is important to note that the dataset also includes instances of all the other events present in the taxonomy to capture the multifaceted soundscape encountered during driving.

The synthetic dataset was meticulously created to simulate various sound events and their combinations, effectively replicating potential distractions encountered in driving scenarios. Isolated sound events (both real and synthetic) were gathered and randomly combined to create 10-second audio mixtures, offering a diverse range of audio samples for training and evaluation purposes. The isolated audio files used in the creation of the synthetic dataset originate from various sources with unrestricted usage rights. Specifically, these sources are: Musical genre classification of audio signals [35], UrbanSound [23], Audio Set [36], Old phone ringtones as midi [37], ASR-CabNois [38], Freesound [39], A singing voice [40], DCASE 2020 Task 4 dataset [41], and A cry corpus features dataset [42]. Detailed information regarding the origins of each type of event can be found in the accompanying Fig. 2, providing a comprehensive reference list.

The collected data underwent a thorough manual review process to ensure authenticity. This involved verifying the existence of the mentioned events and ensuring a uniform quality of the isolated audio samples. Recordings with excessive noise were eliminated, and precise trimming was performed to remove sections devoid of the targeted events. To create the synthetic dataset, Scaper [43] was employed, enabling automated creation of the audio mixtures with precise timing of event insertions. Augmentation techniques such as time stretch or pitch shifting were also utilized to enhance diversity and help generalization.

All the generated mixtures in the dataset contain background noise extracted from real recordings in vehicles [38,39]. The procedure included the random extraction of noise excerpts from the original sources, potentially representing different vehicles and road surfaces. Events were then positioned by Scaper, employing random gains within a specified signal-to-noise ratio (SNR) range of [-20, -10]. This range was chosen to achieve a well-balanced relationship between noise and events, taking into account the loudness considerations inherent in the

Table 1
Number of original isolated events considered and final number of event occurrences in all audio files along with total number of labeled (distractors) and unlabeled (other) events. Both isolated and in-audio events in the train, validation and test sets follow the same proportion as with the number of audio files the same proportion as audio files (~ 80-10-10%).

label		isolated events		dataset appearances	
class	*event*				
RingTone		399		2835	
Vibrating		50		2988	
Notifications		50		2862	
Speech		266		3020	
Cry		545		2950	
Physiological	*Sneeze*	170	*53*	8639	*2882*
	Yawn		*51*		*2861*
	Cough		*66*		*2896*
Pets	*Dog*	470	*295*	5843	*2986*
	Cat		*175*		*2857*
Horns	*BoatHorn*	382	*50*	11693	*2922*
	TruckHorn		*50*		*3004*
	TrainHorn		*50*		*2801*
	CarHorn		*232*		*2966*
Siren	*Police*	150	*50*	8858	*2957*
	Ambulance		*50*		*2991*
	Firefighters		*50*		*2910*
Total no. of distractors		2482		49688	
Other event types		5202		16597	
Total		**7684**		**66285**	

Scaper processing. Additionally, real-world impulse responses measured within vehicles were employed to simulate in-car events with higher realism, for example, simulating speech coming either from the car's audio system or from the occupants. The employed impulse responses correspond to the acoustic paths from a standard car speaker set to 4 different microphone positions within a car-cabin. The microphones (Schoeps CMC-5 stereo pairs with MK5 capsules in an omnidirectional configuration) were placed in the central part of the car at headrest height. The measurement procedure was standard, using sine sweeps as excitation signals. They can be freely downloaded from [44]. Utilizing the synthetic dataset as the primary data source for training deep learning models offers several advantages. It eliminates the need for labor-intensive manual annotation, as the events and their combinations are automatically generated and labeled. Moreover, synthetic data can be extensively leveraged in the training phase, providing a large and diverse set of labeled audio samples.

The synthetic dataset comprises 19,000 audio files, each with a duration of 10 seconds, resulting in a approximate total duration of 53 hours. Out of these files, 15,000 are allocated for training purposes, while the remaining ones were split up and set aside for validation and testing, each of size 2000. To ensure independence in the evaluation process, isolated events from each type were split also in the same proportion as audio files (roughly 80-10-10%) in such a way that there is no overlap of isolated data between the training, validation and testing subsets. The generated audio mixtures for all three sets were carefully constructed to approximately balance the occurrences of each type of event. This means that classes as `Horns` can be up to 4 times more populated as single-event classes as `Cry`. The exact number of isolated events per event type/class and the final number of occurrences along the dataset is shown in Table 1.

The three subsets presented in this study are accompanied by their respective hard labels. These labels indicate the presence of specific events within an audio and provide information about the event's class, beginning, and end timestamps. The event class designations align with the *Event Level* as outlined in the taxonomy. This standardized information not only serves the specific purpose of detecting distractors in this study but can also be utilized across various other problem domains. To ensure standardization, a consolidated approach was adopted, similar to other related problems, where tags are provided within a single text file. This file presents a comprehensive list of hard labels for all audio files, including essential information such as the audio's name, event type, and onset and offset timestamps for each event. This offers flexibility as the provided information can be used to generate different hard labeling formats to suit specific models or frameworks. While the dataset primarily includes hard labels, it is worth noting that weak labels, indicating only the presence or absence of an event within an audio, can be inferred if required.

The dataset is openly accessible through https://www.kaggle.com/datasets/ccastorena/sound-event-detection-for-driver-safety. Researchers and practitioners can utilize this data by simply citing this work as the source.

4. SED models

SED is a crucial audio processing task that aims to automatically recognize and classify various sound events while accurately determining their temporal locations. SED plays a vital role in a wide range of applications, including security systems, human activity analysis, and environmental monitoring. Deep learning neural networks have emerged as a powerful approach for tackling this task effectively. In this section, we discuss two relevant methods commonly applied in SED, particularly in environments similar to the one proposed in this work. Additionally, we explore the utilization of a YOLO-based strategy for the SED task.

4.1. CRNN model

Convolutional recurrent neural network (CRNN) models are known to be a powerful architecture commonly used in SED [22,45,46]. In a CRNN model, the convolutional layers play a crucial role in extracting meaningful features from the input spectrogram. These layers apply a set of learnable filters to the spectrogram, identifying relevant patterns and structures that may represent different sound events. The output of the convolutional layers is a feature map that preserves the time-frequency relationships of the input spectrogram. The recurrent layers in the CRNN model process the temporal sequence of feature activations obtained from the convolutional layers. This allows the model to capture the temporal dependencies and context present in audio signals. By incorporating recurrent connections, such as long short-term memory (LSTM) or gated recurrent unit (GRU), the CRNN model can effectively model long-term dependencies and capture the temporal dynamics of sound events. Finally, dense layers are typically employed as classifiers, taking the output from the recurrent layers and mapping it to class probabilities, indicating the likelihood of each sound event class. The CRNN model has emerged as one of the most widely utilized architectures in the SED field. It has been extensively employed as a baseline in numerous studies and benchmarking initiatives, including the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenges [47,48,22,49,50]. The popularity of the CRNN model in the SED literature underscores its effectiveness and robustness in addressing the complex task of recognizing and classifying sound events in various acoustic environments.

In our study, we will focus on reproducing the baseline architecture [49] of the DCASE 2021-2023 edition, considering the specific variations that exist within this model when applied to application-specific problems. Increasing the number of convolutional or recurrent layers and their sizes would lead to a more powerful predictor but with more overfitting problems. While decreasing them will produce an easier to train but perhaps too simplistic predictor. We opted to maintain the original DCASE 2023 CRNN baseline configuration to facilitate standardized comparisons within the well established and recognized DCASE framework. Our aim is to provide a comprehensive analysis of the CRNN model's capabilities and limitations within the context of synthetic data for SED in a driving scenario.

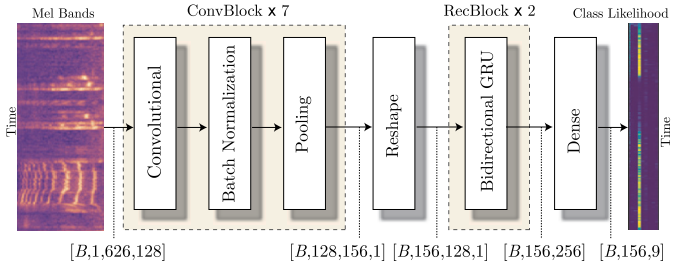


Fig. 3. CRNN model architecture. Brackets indicate input/output dimensions.

4.1.1. Architecture

Fig. 3 describes the architecture and output shapes of the complete CRNN model. The model consists of several convolutional blocks, each comprising a 2D convolutional layer with a kernel size of 3×3. The number of filters for each block is as follows: 16, 32, 64, 128, 128, 128, 128. Additionally, each block includes batch normalization, GLU activation function and pooling operations with specific dimensions of 2×2, 2×2, 1×2, 1×2, 1×2, 1×2, respectively. The initial input undergoes these pooling operations, resulting in a reduction of its time dimension by a factor of 4 and its frequency dimension by a factor of 128. Consequently, the output shape of the convolutional section becomes [B, 128, 156, 1], where B denotes the mini-batch size. To ensure compatibility with the input of the bidirectional recurrent layers, the dimensions of filters and time are transposed, and the excess dimension is eliminated. The bidirectional GRU layer produces an output shape of [B, 156, 256], with 128 hidden units for one direction and 128 for the other. Finally, the dense layer generates the final output, consisting of 156 frames and C different event activations.

The input to the CRNN model is a Mel spectrogram representation of the audio mixtures in the dataset. The Mel spectrum is computed using 128 Mel bands, a window length and FFT size of 2048 points, a Hamming windowing function, and a hop size of 256. With a sampling rate of 16,000 Hz, the Mel spectrogram covers frequencies from 0 Hz to 8,000 Hz, providing a comprehensive representation of the audio's frequency content across different time frames and Mel frequency bands. This spectrogram, with dimensions of 626 frames, 128 Mel bands, and one channel, serves as the input for the CRNN model. Regarding the labeling, the CRNN model employs a frame-by-frame hard label format, where time is segmented into discrete frames. Each frame is assigned a value to indicate the presence or absence of an event, with a value of 1 for frames containing the event and 0 for inactive frames. This labeling format results in a matrix with dimensions corresponding to the number of frames and the number of distinct events to be detected. In this case, the model's output requires 156 frames to align with the selection of 9 events. Alongisde the hard labels, weak tags are automatically derived by identifying presence or absence of a given event.

During the training process, the CRNN model undergoes several important steps to optimize its performance. The training data is augmented using the mix-up technique. Mix-up is a data augmentation technique widely used in machine listening and computer vision [51]. It combines two random training examples by linearly interpolating both data (spectrograms) and labels, creating synthetic samples. This augmentation strategy helps increase the diversity of the training samples, preventing overfitting. Additional augmentation techniques based on mosaicing (introduced in Section 4.2), were applied in conjunction with the mix-up technique for comprehensive and fair evaluation. However, none of the combined augmentation strategies yielded improved results compared to using mix-up alone, which is the only option reported for the CRNN architecture in this paper.

The model is trained for 200 epochs, employing a mini-batch size of 48 and utilizing the Adam optimizer. It initiated with a learning rate of 0.001 and incorporated exponential warmup [52]. The selection process involved identifying the model that exhibited the highest performance on the validation set. The optimization considers three distinct types of errors. Firstly, the hard error is computed to evaluate the event classification accuracy. This is achieved by comparing the model's predictions with the hard labels using the binary cross-entropy (BCE) error metric. The hard error provides insights into how accurately the model can detect different audio events. Secondly, the weak error is calculated by generating a unique prediction vector for all output frames that gets activated (for each class) with the presence of the corresponding class at any output frame. This error metric also utilizes the BCE loss function and enables the model to learn to detect events even in cases where their precise temporal locations are not provided. The weak error evaluation helps evaluate the model's ability to detect and identify relevant events in a more general sense. Lastly, the consistency error is utilized specifically for unlabeled data. This error metric quantifies the mean square error (MSE) between the predictions of the network (Student) and a copy of itself (Teacher) that is updated using an exponential moving average (EMA) function. This error calculation aims to enforce consistency and stability in the model's predictions, even in the absence of labeled data [53].

4.2. YOLO-based model

In this work, we introduce the utilization of the YOLO architecture for SED, which represents a novel approach in the field. While previous works in SED have explored the use of YOLO-based architectures [54,55], they have predominantly focused on earlier versions of YOLO. Our proposed model builds upon the foundation of YOLO, specifically YOLOv5 [24], a popular model initially designed for object detection in images. However, we extend its application to the domain of audio analysis for SED.

The adaptation of a YOLO-like architecture for SED (YOHO) was first proposed in [54], where a version specifically tailored for audio analysis was introduced. Our work draws inspiration from the above approach, which shares the objective of predicting the onset and offset of sound events, aligning with the original bounding box concept in YOLO. However, our current implementation of the YOLO-based model diverges from YOHO in certain aspects. Instead of directly predicting the exact start and end times of sound events, our model estimates the duration, midpoint, and employs different connections in the convolutional layers. Additionally, we incorporate a distinct section for event classification. These adaptations leverage the inherent similarities between image and audio data, enabling the YOLO model to effectively learn and recognize distinctive patterns within spectrograms, facilitating accurate predictions for class activities. While the challenges of SED differ from those of object detection, the YOLO architecture has garnered significant attention and demonstrated impressive performance in various domains. Notably, it stands out for its efficient prediction times and computational scalability, making it an attractive choice for our specific purpose. By leveraging the efficiency of the YOLO architecture in the context of SED [55,56], we aim to develop models that offer accurate detection capabilities with low computational costs. This advancement has far-reaching implications, enabling real-time or resource-constrained applications to benefit from efficient and effective SED models.

In contrast to the CRNN model, our proposed YOLO-based approach does not include recurrent layers for extracting sequential knowledge. Instead, we rely primarily on convolutional networks, adhering to the configuration initially designed for image analysis. This decision prioritizes the efficient utilization of convolutional networks, leveraging their inherent parallel processing capabilities, which can be particularly advantageous in scenarios where computational efficiency and real-time processing are critical considerations.

4.2.1. Architecture

The YOLOv5 architecture consists of three main sections. First, the *backbone* which is a sequence of convolutional blocks that extracts features at different semantic levels. Second, the *neck* that applies a specific

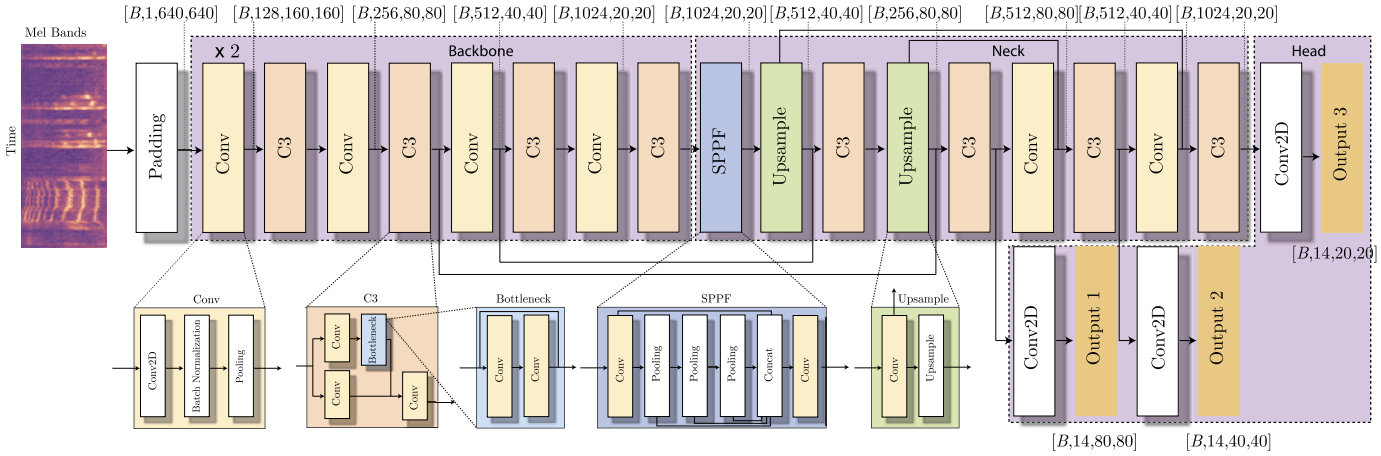


Fig. 4. YOLOv5s model architecture. Brackets indicate input/output dimensions.

version of spatial pyramidal pooling (SPPF) to these feature maps. And the *head* that finally predicts the bounding boxes of objects/events using the feature pyramid constructed by the neck. Different YOLO versions differ in how these three sections are implemented. As in other studies, we have selected YOLOv5 as a very good compromise between performance, adaptability, ease to use and availability [56,57]. From the different YOLOv5 submodels available (with increasing number of filters, pooling layers and depth) we consider only the three smallest ones which are referred to as nano, small and medium, respectively. A graphical representation of the YOLOv5s architecture is shown in Fig. 4.

The input for the YOLO model remains identical to that of the CRNN model, but it is properly padded to match the standard YOLO input shape of 640×640. From the feature pyramid, the YOLOv5 head produces a multi-scale output consisting of three square grids of consecutive resolutions (strides 8, 16 and 32) where each cell contains a bounding box prediction of size $5 + C = 14$ consisting of normalized values of activity (one value), position and size (two values each), and (one-hot encoded) class prediction (one value per class). This allows the detection of events of varying sizes at three different resolution levels. The model processes the pyramid to produce bounding box predictions at each level [24]. Non-maximum suppression [58] is then applied to filter and retain the most confident, non-overlapping boxes, regardless of the resolution that generated them.

Three types of losses at the three resolution levels are considered. Firstly, the location error is computed using the complete intersection over union (CIoU) metric, which evaluates the accuracy of event localization and duration. Secondly, the classification error is measured using binary cross-entropy (BCE), ensuring that events are correctly assigned to their respective classes. Lastly, the objectness error, also calculated with BCE, assesses the confidence level of the model's predictions.

Similar to the experiments conducted with the CRNN, the training of these models entails a 200-epoch training cycle with a batch size of 48 samples, using as well the same optimizer parameters. During training, the model with the best validation performance is selected and retained for further analysis.

During training, mosaic augmentation [59] was employed, which has been shown to be a valuable technique for improving the performance and robustness of YOLO-based object detection models. This method entails the integration of separate training images into a single, enlarged canvas, which is then employed for model training. These individual training images are randomly positioned within the mosaic, and adjustments are made to the bounding box annotations for objects within each image to align them with their new mosaic positions. This approach notably enhances the contextual awareness of the model. The results reported in the next section for YOLO models correspond to the use of both mix-up and mosaic augmentation, even though different combinations of these yield very similar results. Notably, only when mosaicing was completely suppressed the results exhibited a moderate degradation.

5. Performance assessment

In this work, two distinct strategies have been employed to comprehensively evaluate the performance of each predictor. Firstly, event-based detection where each event is assessed as a whole, considering misclassification and the accuracy of predicting onset and offset timestamps. Secondly, segment-based evaluation that is performed by considering fixed-size prediction segments, treating each one as C binary classification problems. Additionally, all models rely on an activity threshold (θ) to generate predictions. Hence, evaluation can be carried out in a threshold independent way, across the entire range of threshold values (global), or locally, by selecting an appropriate (or ideally the best) threshold value for assessment.

For evaluating global, event-based performance, we adopted the widely accepted polyphonic sound event detection scores (PSDS) provided by the `psds_eval` toolbox [60], a recognized reference in the DCASE community. The computation of PSDS involves deriving an ROC (receiver operating characteristic) curve for all events within a given audio, encompassing 50 operating points. This process assigns true/false positives (TP/FP) based on the detection tolerance criterion (DTC), the ground truth intersection criterion (GTC), and other relevant parameters. In our analysis, we explored two distinct sets of criteria, yielding two different metrics: PSDS1, where $DTC = GTC = 0.7$, in line with common practices in general SED tasks; and PSDS0, where $DTC = GTC = 0.5$, leading to heightened TP rates and enabling the exploration of an alternative curve and operational region within the threshold spectrum.

When evaluating segment-based metrics, a shift in perspective occurs compared to the event-based approach. In this context, TP correspond to frames where event activations are correctly recognized. FP encompass frame misactivations, onset anticipations, and offset delays, while false negatives (FN) gauge undetected active frames, offset advancements, and onset delays. Additionally, true negatives (TN) correspond to correctly identified non-active frames. All segment-based metrics considered here use output prediction segments of one tenth of a second which leads to 100 output frames for inputs of 10 seconds. The segment-based metrics are then computed through (micro-)averaging by identifying TP, FP, TN and FN for each class.

The averaged area under the curve (AUC) value gives us a measure that takes into account the whole range of values of the threshold θ. Apart from this, the AUC_{FP} value that restricts the above area for FP rates below a given value has been considered. This measure is more representative because only predictors that lead to low FP rates are of

Table 2
Performance measures obtained over the test partition. All local measures have been computed using the best possible threshold. Best performing model is marked as bold and squared boxed figures in the CRNN column means that CRNN performs better than at least one of the YOLO models.

		CRNN	YOLOv5n	YOLOv5s	YOLOv5m
event-based local global	PSDS1	0.24	0.46	0.46	**0.47**
	PSDS0	0.37	0.56	0.57	**0.59**
	$F_1(1)$	0.38	**0.60**	0.59	**0.60**
	$F_1(0)$	0.50	0.68	0.67	**0.69**
segment-based local global	AUC	**0.92**	0.85	0.81	0.81
	$AUC_{0.1}$	0.68	**0.78**	0.77	0.77
	F_1	0.65	0.75	**0.76**	**0.76**
	Acc	0.95	0.96	0.96	**0.97**
	G-mean	0.80	**0.84**	**0.84**	0.83

interest in practice.[1] A maximum FP rate of 0.1 has been considered in this work, i.e. $AUC_{0.1}$.

The F_1 score is widely used and important because it offers a balanced evaluation of classification models. It corresponds to the harmonic mean of precision (ratio of TP over all predicted as positive) and recall (or TP rate). This balance is particularly valuable in SED scenarios, where the data imbalance problem resulting from sparse event activations carries different consequences (i.e. false alarms versus missed events). As a single metric, the F_1 score provides a clear and concise measure of the overall performance of a model in capturing relevant instances while minimizing errors. In addition to this, other classification metrics such as the overall accuracy (Acc), and the geometric mean of TP and TN rates (G-mean) have also been included as local, segment-based performance measures.

In the local, event-based scenario, it is important to note that traditional classification measures cannot be directly calculated due to the lack of a meaningful concept of TN. However, by utilizing the TP, FP, and FN derived from the computation of PSDS, it becomes possible to derive precision and recall scores. As a result, we incorporate the class-averaged $F_1(x)$ measures for both PSDS metrics, where x takes values of 0 and 1, respectively. These measures correspond to the optimal operating points on the respective exploration curves, as elaborated in [61].

6. Results and discussion

Table 2 provides a concise overview of the performance measures obtained for the test dataset using the four prediction models explored in this study. Notably, the CRNN model consistently demonstrates inferior performance across different metrics, except for AUC, when compared to the other models.

This observation can be better understood by looking at Fig. 5, which shows the four ROC curves along with the corresponding best predictors (according to F_1). Overall, all YOLO models exhibit saturated ROC curves, meaning that the TP rate remains constant even when very low θ values are used. This behavior arises because YOLO models tend to provide very small output values for some classes (below the lowest θ value explored), which leads to many non-detected TP. As a result, the AUC measure is highly influenced by such saturation effect. The $AUC_{0.1}$ measure leads to values that are highly correlated with all other measures including the global, event-based ones as they correspond to almost all the operating range of the threshold θ. We can grasp this operating region from a different view if we observe Fig. 6 where the segment-based F_1-score across the whole threshold range has been plotted for all predictors, along with the same best individual predictors for each model. Note that in general, all predictors will tend to obtain zero F_1 scores as the threshold is decreased but at different rates. Apart from

[1] In fact, PSDS also applies this kind of restriction but in terms of number of FP per hour of audio.

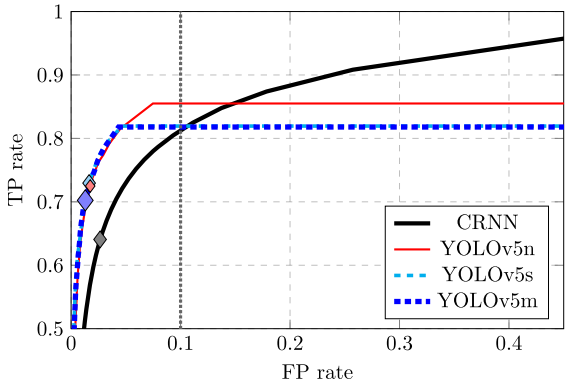


Fig. 5. ROC curves corresponding to the different models considered on the test partition. Marked points on the curves correspond to the best threshold values according to the segment-based F_1-score.

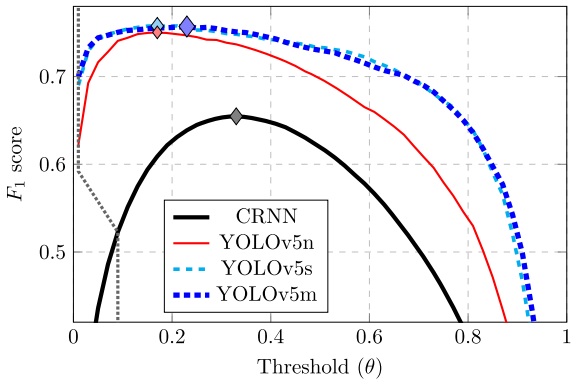
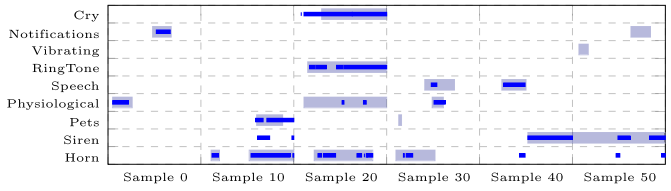


Fig. 6. Averaged F_1 scores for different threshold values corresponding to the test partition. Best values and corresponding thresholds are marked on the curves. The region at the left of the dotted line corresponds to FP rates higher than 0.1.

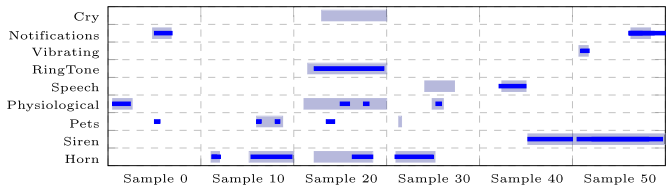
that, the fact that YOLO detects bounding boxes instead of generating a mask of predictions as CRNN gives as a result that in practice we very hardly obtain FP values above 0.1 even for very low threshold values as can be observed in the figure.

The event-based performance measures presented in Table 2 exhibit the same general trend. However, the global indicators display relatively lower values, particularly noticeable in the case of PSDS1. This value can be readily compared to outcomes from other studies focused on general SED problems [62,61]. This discrepancy arises due to the highly restrictive nature of the PSDS family of scores. These scores aim to discern among predictors that perform exceptionally well on a given task, a level of achievement that is still beyond our current reach within our specific context. Furthermore, these metrics strongly penalize false positives, even when originating from intermittent events that hold significant importance in our dataset. When assessing specific predictors with fixed thresholds, mirroring real-world scenarios, we attain more realistic results in the corresponding event-based F_1(x)-scores, as computed through the `sed_eval` toolbox.

In fact, the behavior of all four models for both $F_1(x)$-scores closely mirrors that of the segment-based F_1 in Fig. 6, showcasing remarkably similar threshold values for the most effective predictors. In general, the performance outcomes of YOLO models demonstrate a notable stability across various threshold values. The optimal θ values consistently hover around 0.2, but very good results are also achievable within a range spanning from 0.1 to 0.5 or beyond. In contrast, the CRNN model exhibits an optimal value at approximately 0.4 (0.47 based on the $F_1(0)$-score) and its performance rapidly deteriorates as this value is either increased or decreased.



(a) CRNN (θ=0.33)



(b) YOLOv5n (θ=0.17)

Fig. 7. Prediction outputs filtered at given thresholds θ (shown in blue) for six particular audio segments from the test set. Ground-truth is displayed in light blue.

Table 3
Number of parameters, compiled model size, and GPU/CPU prediction times (RTX3060ti/Cortex A72) for the SED models considered in this work.

	CRNN	YOLOv5n	YOLOv5s	YOLOv5m
params (M)	1.1	1.9	7.2	20
model size (MB)	4.5	10	36.5	100
GPU time (ms)	34	29	33	35
CPU time (ms)	292	226	387	731

To visually illustrate the predictive outcomes of CRNN and YOLO models, we present specific prediction results for 10-second audio inputs from the test set in Fig. 7. Upon analyzing the graphical representation, it becomes evident that YOLOv5n consistently outperforms CRNN. YOLOv5n (and the other YOLO models, which provide similar predictions) generates fewer false positives and misses fewer events compared to CRNN. This example emphasizes the robustness of YOLO-based models and their potential for SED.

To provide further insight into model complexities and their impact on approximate prediction times, Table 3 provides the number of parameters, the compiled model size (utilizing the standard Open Neural Network Exchange [63]), and the average inference time for generating predictions within a 10-second audio clip. Assessment on an RTX3060ti GPU and a Cortex A72 (Raspberry Pi 4) CPU emphasizes the superiority of the basic YOLOv5n model in balancing performance and complexity.

6.1. Some preliminary results with real data

Although generating a fully annotated audio dataset based on real driving scenarios is beyond the scope of this study, it remains crucial to assess the performance of models trained on our synthetic dataset when applied to actual real-world data. For this purpose, we conducted a recording during a real car trip, capturing instances of select considered events that occurred naturally or were deliberately provoked, aimed at testing the trained models. This resulting audio, while relatively concise, underwent meticulous annotation and segmentation into 10-second intervals. This collection comprises 685 audio segments, about a quarter the size of our synthetic dataset, with annotations for five out of the nine distractor events (`Speech`, `Horn`, `Physiological`, `RingTone`, and `Notification`). It is noteworthy that not only do the sound context and noise conditions differ significantly, but the frequency and proportion of active/inactive frames for each event type also differ considerably from those observed in the synthetic audio files. Results obtained from this real data should be interpreted with caution, considering the inherent limitations of this preliminary experiments, including the limited size and diversity of the dataset, which may impact the generalizability of findings beyond the specific model comparisons.

Table 4
Performance measures obtained over the real test set. All local measures have been computed using the best possible threshold. Best performing model is marked as bold and squared boxed figures in the CRNN column means that CRNN performs better than at least one of the YOLO models.

		CRNN	YOLOv5n	YOLOv5s	YOLOv5m
event-based	PSDS1	0.01	0.01	0.03	**0.04**
global	PSDS0	0.04	0.01	0.03	**0.10**
local	$F_1(1)$	0.18	0.40	**0.44**	0.39
	$F_1(0)$	0.25	**0.52**	0.44	0.39
segment-based	AUC	0.87	**0.92**	0.89	0.87
global	$AUC_{0.1}$	0.51	0.71	**0.78**	0.73
local	F_1	0.59	0.71	**0.77**	0.73
	Acc	0.93	0.94	**0.95**	0.94
	G-mean	0.79	0.89	**0.90**	0.89

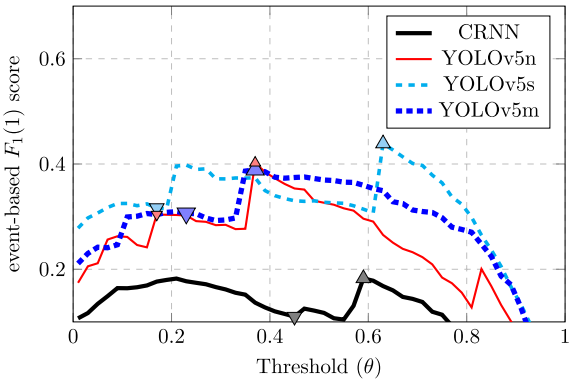
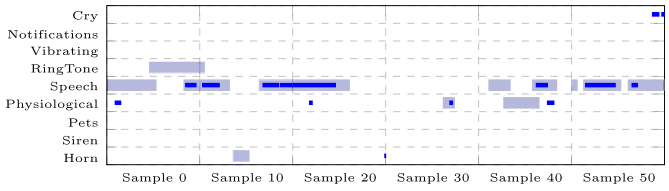


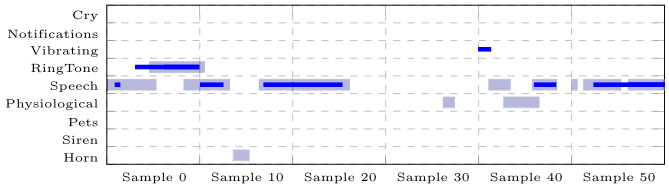
Fig. 8. Averaged $F_1(1)$ scores for different threshold values corresponding to the real test set. Best values and corresponding thresholds are marked on the curves (triangles up). The best predictors for synthetic data are also marked (triangles down).

The evaluation process employed to generate the previous (synthetic) results was also applied to the real-world test data, and the corresponding outcomes are depicted in Table 4. As anticipated, the real-world results generally lag behind the synthetic values. Nevertheless, the established pattern persists, with CRNN exhibiting relatively inferior performance compared to the YOLO models. Indeed, even the global AUC measure records the most favorable outcomes for two of the YOLO models.

Drawing valid conclusions from PSDS scores is difficult due to the extremely low values obtained. The infrequent occurrence of events of interest within our real data, coupled with the maximum effective false positive rate implicit in the PSDS definition [60], lead to exceedingly low PSDS values across all models tested. In this context, very low PSDS values indicate that the best operating points of the predictors are associated with a FP rate beyond the limit established by this metric. A closer examination is provided through the $F_1(1)$-scores corresponding to all thresholds, visualized in Fig. 8. It becomes evident that all models experience substantial drops in performance when compared to their synthetic counterparts. Furthermore, the optimal θ values for the best predictors are generally higher. Most importantly, it is apparent that these predictors struggle to accurately detect complete events, especially at lower θ values. This issue is further evident in the case of CRNN, which displays even more erratic and inferior performance across the entire range of θ. These observations can be attributed to the fact that only five event types are genuinely present in the audio recordings, each characterized by statistics significantly divergent from those in the training dataset.



(a) CRNN (θ=0.59)



(b) YOLOv5n (θ=0.37)

Fig. 9. Prediction outputs filtered at given thresholds θ (shown in blue) for six in a row audio segments from the real audio. Ground-truth is displayed in light blue.

Interestingly, in terms of segment-based performance measures, it has been observed that all models, including CRNN, exhibit a similar pattern to that exhibited by synthetic data. This similarity is evident despite a slight downward shift in the values. Low event-based metric values do not necessarily indicate predictor failure. In fact, as depicted in Fig. 9, the predictions of both models are capable of accurately detecting certain events. Generally, CRNN exhibits more false positives and undetected events than YOLOv5n, although the predictions of the latter often extend beyond the ground truth tags. Indeed, YOLOv5n presents fewer false positives, demonstrating a superior capability to distinguish between event and non-event frames. However, its tendency to predict longer event durations might arise from its capacity to capture a broader audio context. Conversely, CRNN's reliance on recurrent layers could impede consistent event tracking, potentially leading to premature event terminations.

7. Conclusions

In this study, we developed a comprehensive taxonomy for driving sounds, which serves as a structured framework for categorizing and understanding the various types of sounds encountered during driving scenarios. By organizing the sounds into different levels, we aimed to provide a systematic approach to identify and analyze recognized auditory distractors that can significantly impact driver safety. To facilitate the training and evaluation of sound event detection systems, we introduced a carefully created dataset that simulates audio recordings in driving scenarios comprising mixtures of commonly occurrying events, eliminating the need for labor-intensive manual annotation. In our approach to sound event detection, we explored two different architectures: a baseline CRNN model and a YOLO-based model adapted for audio analysis. Although both showed promising results, the YOLO-based model significantly outperformed the baseline model in all the analyzed metrics. The taxonomy, dataset, and models presented in this work contribute to the advancement of sound event detection systems tailored for driving safety. By identifying and addressing auditory distractors, we aim to enhance driver awareness and mitigate potential risks on the road.

Future work in this domain could focus on further refining the models, exploring additional data processing techniques, and investigating the integration of multiple sensor inputs to enhance the overall driving safety experience. By continuing to innovate in sound event detection, we can make meaningful strides towards safer and more attentive driving practices for all road users.

CRediT authorship contribution statement

Carlos Castorena: Data curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Maximo Cobos:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing. **Jesus Lopez-Ballester:** Software, Writing – review & editing. **Francesc J. Ferri:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Authors would like to thank the Spanish Research Agency (AEI) and the European Regional Development Fund (ERDF) for partially funding this research within the projects with references TED2021-131003B-C21 and PID2022-137048OB-C41 funded by MCIN/AEI/10.13039/501100011033 and by the "EU Union NextGenerationEU/PRTR". We also thank *Generalitat Valenciana*, the *Santiago Grisolía* program for financing this work (GRISOLIAP/2021/060, CPI-21-232). Finally, the authors acknowledge as well the Artemisa computer resources funded by the EU ERDF and Comunitat Valenciana, and the technical support of IFIC (CSIC-UV).

References

[1] The sustainable development goals report 2022. Available from: https://unstats.un.org/sdgs/report/2022/, July 2022.
[2] Organization WH. Global status report on road safety 2018: summary (no. who/nmh/nvi/18.20). World Health Organization; 2018.
[3] Gupta A, Anpalagan A, Guan L, Khwaja AS. Deep learning for object detection and scene perception in self-driving cars: survey, challenges, and open issues. Array 2021;10:100057. https://doi.org/10.1016/j.array.2021.100057.
[4] Vicente F, Huang Z, Xiong X, De la Torre F, Zhang W, Levi D. Driver gaze tracking and eyes off the road detection system. IEEE Trans Intell Transp Syst 2015;16(4):2014–27. https://doi.org/10.1109/TITS.2015.2396031.
[5] Jiang S, Guo Z, Zhao S, Wang H, Jing W. Ce-gan: a camera image enhancement generative adversarial network for autonomous driving. In: 2022 IEEE 9th international conference on data science and advanced analytics (DSAA). IEEE; 2022. p. 1–6.
[6] Halim Z, Rehan M. On identification of driving-induced stress using electroencephalogram signals: a framework based on wearable safety-critical scheme and machine learning. Inf Fusion 2020;53:66–79. https://doi.org/10.1016/j.inffus.2019.06.006. Available from: https://www.sciencedirect.com/science/article/pii/S1566253518303221.
[7] Faysal K, Sahadev R. Real-time vehicular accident prevention system using deep learning architecture. Expert Syst Appl 2022;206:117837. https://doi.org/10.1016/j.eswa.2022.117837. Available from: https://www.sciencedirect.com/science/article/pii/S0957417422010958.
[8] Eraqi HM, Abouelnaga Y, Saad MH, Moustafa MN. Driver distraction identification with an ensemble of convolutional neural networks. J Adv Transp 2019;2019. https://doi.org/10.1155/2019/4125865.
[9] Srinivasan K, Garg L, Datta D, Alaboudi AA, Jhanjhi N, Agarwal R, et al. Performance comparison of deep CNN models for detecting driver's distraction. CMC Comput Mater Continua 2021;68(3):4109–24.
[10] Li W, Huang J, Xie G, Karray F, Li R. A survey on vision-based driver distraction analysis. J Syst Archit 2021;121:102319. https://doi.org/10.1016/j.sysarc.2021.102319.
[11] Marcillo P, Valdivieso Caraguay AL, Hernandez-Alvarez M. A systematic literature review of learning-based traffic accident prediction models based on heterogeneous sources. Appl Sci 2022;12(9). https://doi.org/10.3390/app12094529. Available from: https://www.mdpi.com/2076-3417/12/9/4529.
[12] Li W, Gkritza K, Albrecht C. The culture of distracted driving: evidence from a public opinion survey in Iowa. Transp Res, Part F Traffic Psychol Behav 2014;26:337–47. https://doi.org/10.1016/j.trf.2014.01.002.

[13] Prat F, Planes M, Gras M, Sullman M. An observational study of driving distractions on urban roads in Spain. Accid Anal Prev 2015;74:8–16. https://doi.org/10.1016/j.aap.2014.10.003.

[14] Prat F, Gras M, Planes M, Font-Mayolas S, Sullman M. Driving distractions: an insight gained from roadside interviews on their prevalence and factors associated with driver distraction. Transp Res, Part F Traffic Psychol Behav 2017;45:194–207. https://doi.org/10.1016/j.trf.2016.12.001.

[15] Farmer CM, Braitman KA, Lund AK. Cell phone use while driving and attributable crash risk. Traffic Inj Prev 2010;11(5):466–70. https://doi.org/10.1080/15389588.2010.494191.

[16] Edwards S, Wundersitz L, et al. Distracted driving: prevalence and motivations. Accid Anal Prev 2019;54:99–107.

[17] Huisingh C, Griffin R, McGwin Jr G. The prevalence of distraction among passenger vehicle drivers: a roadside observational approach. Traffic Inj Prev 2015;16(2):140–6.

[18] Carney C, Harland KK, McGehee DV. Examining teen driver crashes and the prevalence of distraction: recent trends, 2007–2015. J Saf Res 2018;64:21–7. https://doi.org/10.1016/j.jsr.2017.12.014.

[19] Bernstein JJ, Bernstein J. Texting at the light and other forms of device distraction behind the wheel. BMC Public Health 2015;15:1–5. https://doi.org/10.1186/s12889-015-2343-8.

[20] Sullman MJ, Prat F, Tasci DK. A roadside study of observable driver distractions. Traffic Inj Prev 2015;16(6):552–7. https://doi.org/10.1080/15389588.2014.989319.

[21] Cai X, Wang C, Chen S, Lu J. Model development for risk assessment of driving on freeway under rainy weather conditions. PLoS ONE 2016;11(2):e0149442. https://doi.org/10.1371/journal.pone.0149442.

[22] Turpault N, Serizel R, Parag Shah A, Salamon J. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In: Workshop on detection and classification of acoustic scenes and events; 2019.

[23] Salamon J, Jacoby C, Bello JP. A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on multimedia, association for computing machinery; 2014. p. 1041–4.

[24] Jocher G. YOLOv5 SOTA realtime instance segmentation. Available from: https://doi.org/10.5281/zenodo.3908559. https://github.com/ultralytics/{YOLO}v5, May 2020.

[25] Lindborg P. A taxonomy of sound sources in restaurants. Appl Acoust 2016;110:297–310. https://doi.org/10.1016/j.apacoust.2016.03.032. Available from: https://www.sciencedirect.com/science/article/pii/S0003682X16300652.

[26] Bi T, Olugbade T, Mathur A, Holloway C, Singh A, Costanza E, et al. A taxonomy of noise in voice self-reports while running. In: Adjunct proceedings of the 2022 ACM international joint conference on pervasive and ubiquitous computing and the 2022 ACM international symposium on wearable computers, UbiComp/ISWC '22 adjunct. New York, NY, USA: Association for Computing Machinery; 2023. p. 229–32.

[27] Gazder U, Assi KJ. Determining driver perceptions about distractions and modeling their effects on driving behavior at different age groups. J Traffic Transp Eng (Engl Ed) 2022;9(1):33–43. https://doi.org/10.1016/j.jtte.2020.12.005.

[28] Regan MA, Oviedo-Trespalacios O. Driver distraction: mechanisms, evidence, prevention, and mitigation. In: The vision zero handbook: theory, technology and management for a zero casualty policy. Springer; 2022. p. 1–62.

[29] Nagahama A, Tanaka K, Feliciani C, Cui G, Wada T. Effects of urban landscape and soundscape on driving behavior. In: 2022 IEEE conference on cognitive and computational aspects of situation management (CogSIMA); 2022. p. 84–8.

[30] Prohn MJ, Herbig B. Potentially critical driving situations during "blue-light" driving: a video analysis. West J Emerg Med 2023;24(2):348. https://doi.org/10.5811/westjem.2022.8.56114.

[31] Koppel S, Charlton J, Kopinathan C, Taranto D. Are child occupants a significant source of driving distraction? Accid Anal Prev 2011;43(3):1236–44. https://doi.org/10.1016/j.aap.2011.01.005.

[32] Ma Y, Sanchez V, Nikan S, Upadhyay D, Atote B, Guha T. Real-time driver monitoring systems through modality and view analysis. Available from: arXiv:2210.09441, 2022. https://doi.org/10.48550/arXiv.2210.09441.

[33] Bamney A, Megat-Johari N, Kirsch T, Savolainen P. Differences in near-crash risk by types of distraction: a comparison of trends between freeways and two-lane highways using naturalistic driving data. Transp Res Rec 2022;2676(2):407–17. https://doi.org/10.1177/03611981211043817.

[34] Zou T, Guo H, Khaloei M, MacKenzie D, Boyle LN. Examining the relationships between multimodal environments and multitasking driving behaviors. Transp Res Rec 2023;2677(2):944–57. https://doi.org/10.1177/03611981221110223.

[35] Tzanetakis G, Cook P. Musical genre classification of audio signals. IEEE Trans Speech Audio Process 2002;10(5):293–302. https://doi.org/10.1109/TSA.2002.800560.

[36] Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, et al. Audio set: an ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2017. p. 776–80.

[37] Torosyan N. Old phone ringtones as midi [cited 2023]. Available from: https://www.kaggle.com/datasets/narektorosyan/old-phone-ringtones-as-midi, Oct 2019.

[38] ASR-CabNois: a cabin noise dataset [cited 2023]. Available from: https://magichub.com/datasets/in-vehicle-noise-dataset/, 2023.

[39] Freesound [cited 2023]. Available from: https://freesound.org/home/, 2023.

[40] Wilkins J, Seetharaman P, Wahl A, Pardo B. Vocalset: a singing voice dataset. Available from: https://doi.org/10.5281/zenodo.1193957, 2018.

[41] Serizel R, Turpault N, Shah A, Salamon J. Sound event detection in synthetic domestic environments. In: ICASSP 2020 - 45th international conference on acoustics, speech, and signal processing; 2020. p. 86–90. Available from: https://hal.inria.fr/hal-02355573.

[42] Valani, donate-a-cry-corpus-features-dataset [cited 2023]. Available from: https://www.kaggle.com/datasets/bhoomikavalani/donateacrycorpusfeaturesdataset, 2023.

[43] Salamon J, MacConnell D, Cartwright M, Li P, Bello JP. Scaper: a library for soundscape synthesis and augmentation. In: 2017 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA). IEEE; 2017. p. 344–8.

[44] van Saane F. Impulse responses [cited 2023]. Available from: https://fokkie.home.xs4all.nl/IR.htm, 2004.

[45] Martín-Morató I, Harju M, Ahokas P, Mesaros A. Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation. In: Proc. IEEE int. conf. acoustic., speech and signal process. ICASSP; 2023. p. 902–14.

[46] Spoorthy V, Shashidhar GK. Polyphonic sound event detection using Mel-pseudo constant q-transform and deep neural network. IETE J Res 2023;0(0):1–13. https://doi.org/10.1080/03772063.2023.2253768.

[47] Serizel R, Turpault N, Eghbal-Zadeh H, Shah AP. Large-scale weakly labeled semi-supervised sound event detection in domestic environments. In: Proceedings of the detection and classification of acoustic scenes and events 2018 workshop (DCASE2018); 2018. p. 19–23. Available from: https://hal.inria.fr/hal-01850270.

[48] Adavanne S, Politis A, Nikunen J, Virtanen T. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE J Sel Top Signal Process 2018;13(1):34–48. https://doi.org/10.1109/JSTSP.2018.2885636.

[49] Ronchini F, Serizel R, Turpault N, Cornell S. The impact of non-target events in synthetic soundscapes for sound event detection. In: Proceedings of the 6th detection and classification of acoustic scenes and events 2021 workshop (DCASE2021); 2021. p. 115–9.

[50] IEEE-AASP. DCASE 2023 challenge website [cited 2023]. Available from: https://dcase.community/challenge2023/index, 2023.

[51] Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: beyond empirical risk minimization. In: International conference on learning representations; 2018.

[52] Ma J, Yarats D. On the adequacy of untuned warmup for adaptive optimization. In: Proceedings of the AAAI conference on artificial intelligence, vol. 35. 2021. p. 8828–36.

[53] Tarvainen A, Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. Available from: arXiv:1703.01780, 2018.

[54] Venkatesh S, Moffat D, Miranda ER. You only hear once: a YOLO-like algorithm for audio segmentation and sound event detection. Appl Sci 2022;12(7):3293.

[55] Tiwari S, Lakhotia K, Mulimani M. Evaluating robustness of you only hear once (YOHO) algorithm on noisy audios in the voice dataset. In: 35th conference on neural information processing systems (NeurIPS 2021); 2021.

[56] Zheng J-C, Sun S-D, Zhao S-J. Fast ship detection based on lightweight YOLOv5 network. IET Image Process 2022;16(6):1585–93.

[57] Dong X, Yan S, Duan C. A lightweight vehicles detection network model based on YOLOv5. Eng Appl Artif Intell 2022;113:104914.

[58] Liu L, Zhou L. Inverted non-maximum suppression for more accurate and neater face detection. Available from: arXiv:2305.10593, 2023.

[59] Bochkovskiy A, Wang C-Y, Liao H-YM. Yolov4: optimal speed and accuracy of object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2020. p. 163–78.

[60] Bilen C, Ferroni G, Tuveri F, Azcarreta J, Krstulovic S. A framework for the robust evaluation of sound event detection. Available from: arXiv:1910.08440, 2020.

[61] Ferroni G, Turpault N, Azcarreta J, Tuveri F, Serizel R, Bilen C, et al. Improving sound event detection metrics: insights from DCASE 2020. In: ICASSP 2021 - 2021 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2021. p. 631–5.

[62] Mesaros A, Heittola T, Virtanen T. Metrics for polyphonic sound event detection. Appl Sci 2016;6(6):162. https://doi.org/10.3390/app6060162.

[63] Bai J, Lu F, Zhang K. Onnx: open neural network exchange github (online), [cited 2023]. Available from: https://github.com/onnx/onnx, 2023. [Accessed 1 September 2023].