DOI: 10.61782/fa.2025.0006

FORUM ACUSTICUM
EURONOISE 2025

A COMPARATIVE ANALYSIS OF INTERIOR VS. EXTERIOR VEHICLE MICROPHONE PLACEMENT FOR ACOUSTIC EVENT DETECTION IN DRIVING ENVIRONMENTS

Carlos Castorena[1*] **Lucas Banchero**[2] **Juan A. De Rus**[1] **Francisco Vacalebri**[2]
Sandra Roger[1] **Jose M. Mossi**[2] **Jose J. López**[2] **Francesc J. Ferri**[1] **Maximo Cobos**[1]
[1] Departament d'Informàtica, Universitat de València, Burjassot, Spain
[2] Institute of Telecommunications and Multimedia Applications,
Universitat Politecnica de Valencia, Valencia, Spain

ABSTRACT

Safe driving depends on both internal and external factors of the vehicle, including those manifested as acoustic signals. Sounds, whether external, such as sirens or horns, or internal, such as conversations between passengers or the sound system, provide critical information to identify events that could compromise safety. The placement of microphones used for monitoring and feeding into an artificial intelligence-based detection system plays a crucial role. Microphones placed externally are essential for capturing sounds like sirens or horns, but they face challenges such as wind noise and vibrations caused by the movement of the vehicle. On the other hand, detecting these external events from the interior presents difficulties due to attenuation or distortion caused by the acoustic insulation of the body of the vehicle. This work explores the relevance of microphone placement by comparing the performance of models when processing data captured separately from the interior and exterior of vehicles. The challenges associated with capture are also discussed.

Keywords: *sound detection, mic placement, road safety*

1. INTRODUCTION

Audio-based event detection systems offer distinct advantages over vision-based systems in vehicle safety and driver assistance, particularly in scenarios where a direct line of sight to the event source is obstructed [1]. The ability to recognize critical acoustic cues, such as emergency vehicle sirens or horn sounds, enhances situational awareness and supports timely decision-making while driving [2, 3]. However, the performance of such systems is highly sensitive to microphone placement and their robustness in dynamic, noisy environments.

Microphone positioning plays a critical role in determining the accuracy and reliability of audio-based detection systems. External microphones, mounted outside the vehicle, are well-suited to capturing environmental sounds, such as approaching emergency vehicles or honking cars. However, they are vulnerable to various interferences, including wind noise, road noise, and other environmental disturbances. To mitigate these issues, approaches such as physical filters [4], digital signal processing techniques [5], and deep learning-based noise reduction methods [6] have been proposed. While these methods enhance detection performance, they require careful optimization to remain effective under diverse real-world conditions.

In contrast, placing microphones inside the vehicle provides a more controlled acoustic environment, thereby reducing exposure to external noise. Nonetheless, this setup introduces its own challenges, such as sound attenuation caused by the vehicle's structure and interference from internal noise sources like conversations, infotainment systems, and engine vibrations. Additionally, capturing external acoustic events from within the cabin is complicated by the sound-insulating properties of modern vehicle designs. To address these challenges, advanced

*Corresponding author: carlos.castorena@uv.es.

deep learning models capable of processing polyphonic sound events [7, 8]—where multiple overlapping sounds occur simultaneously—are increasingly employed. These models improve the system's ability to differentiate between relevant and irrelevant auditory signals within the complex acoustic environment of a vehicle interior.

Recent research has also explored multimodal approaches [9] that integrate both audio and visual data to improve event detection capabilities. By combining acoustic signals with visual cues, these systems can compensate for limitations in either modality, leading to improved overall detection accuracy. However, multimodal systems introduce additional complexities, including higher computational demands, synchronization requirements between audio and video streams, and sensitivity to adverse environmental conditions (e.g., poor lighting or occlusions) that may impair one or both modalities. These limitations highlight the need for thoughtful design and implementation when considering multimodal solutions.

This study evaluates the impact of microphone placement—inside versus outside the vehicle—on the performance of three state-of-the-art deep learning models: the Audio Spectrogram Transformer (AST) [4, 10], Convolutional Recurrent Neural Networks (CRNN) [11, 12], and a version of YOLO adapted for audio spectrogram inputs [8, 13]. Rather than focusing solely on performance comparisons, the primary objective is to investigate how microphone placement, ambient noise conditions, and overlapping sound events influence the ability of each model to detect key auditory events, such as sirens and horn sounds, under realistic driving scenarios. By examining detection accuracy across different microphone configurations, this work sheds light on the trade-offs inherent in each setup and contributes to the development of more robust and context-aware audio-based detection systems for intelligent vehicle applications.

2. METHODOLOGY

2.1 Data Collection

To evaluate model performance under different acoustic conditions, two distinct sets of audio recordings were collected. Recordings were made using WM-61A DIY and XYH-6 microphones connected to a Zoom H6 recorder to ensure synchronized multi-channel capture. Two channels were placed inside the vehicle, centrally positioned, while four were mounted externally—two on each side. The external microphones were installed on a stable base and fitted with physical wind filters, which have been shown in prior studies to effectively reduce wind interference. The microphone layout is shown in Figure 1.

In Scenario 1, recordings were conducted both inside and outside the vehicle under low-noise conditions . No additional sound sources from the inside—such as radio, mobile phones, or conversations—were present . This scenario represents an idealized environment with minimal internal interference. A total of 1245 seconds were recorded, which were segmented into 4-second clips using a 1-second sliding window, resulting in 23 horn and 94 siren events.

Scenario 2 was designed to simulate a realistic and noisy in-cabin environment. In this case, background sounds including conversations, radio playback, and mobile phone usage were deliberately introduced. The total duration of this recording was 877 seconds, yielding 29 horn and 179 siren events after segmentation using the same process as in Scenario 1.

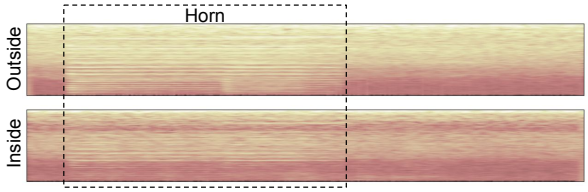
Both datasets were segmented into overlapping 4-second clips using a 1-second stride to construct a continuous and structured dataset. Rather than adopting a traditional sound event detection (SED) framework, each segment was assigned a single label—*Horn*, *Siren*, or *No Event*—based on the presence of a target sound within the

Figure 1: Microphone placement in the vehicle. Microphones M1–M4 (WM-61A DIY) are positioned externally, while microphone M5 (XYH-6, two channels) is located inside.

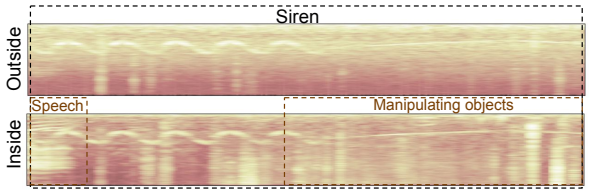
clip. This simplified labeling strategy was chosen to facilitate a direct and controlled comparison of model performance across different microphone placements, focusing on the detectability of key acoustic events rather than precise time event localization or boundary detection. While this approach does not capture the full temporal dynamics of sound events, it provides a practical and interpretable means of evaluating classification accuracy under varying acoustic conditions.

All recordings were conducted with the vehicle in a stationary state to minimize wind-induced noise, whose limited impact has been demonstrated in [4]. The recording site was selected to ensure consistent exposure to traffic-related sounds, resulting in a naturally high frequency of siren and horn events.

To illustrate the acoustic differences between scenarios, Figure 2 shows representative 4-second audio segments from both conditions: (a) a clean sample from Scenario 1, and (b) a more complex recording from Scenario 2, where overlapping internal sounds such as speech and object interaction are present.



(a) Scenario 1: 4-second audio segment with minimal background noise.



(b) Scenario 2: 4-second audio segment with overlapping in-cabin events (e.g., speech, object handling).

Figure 2: Representative audio samples from the two recording scenarios. Scenario 1 provides a low-noise baseline, while Scenario 2 reflects more realistic in-vehicle acoustic conditions.

2.2 Detection Models

To investigate how microphone placement affects the detection of relevant acoustic events in vehicular contexts, we implemented three deep learning-based models with distinct architectural characteristics: the Audio Spectrogram Transformer (AST), a Convolutional Recurrent Neural Network (CRNN), and a fully convolutional architecture based on YOLO. While the primary goal is not to compare absolute model performance, these architectures offer complementary perspectives on how different microphone configurations influence the perception and classification of acoustic signals. Each model was adapted to produce standardized outputs—three class probabilities for *Siren*, *Horn*, and *No Event*—to enable a consistent evaluation framework.

The AST model [14] leverages a Multihead Self-Attention mechanism to learn long-range dependencies within audio sequences. This capacity to capture global contextual information makes it particularly effective for identifying complex acoustic patterns, even in the presence of noise or temporal distortion. In this study, AST was trained using 4-second Mel spectrograms with 128 Mel bands, a 32 ms analysis window, and a 10 ms hop size. Unlike the other models, AST processes stereo audio by jointly analyzing two input channels. For in-cabin detection, the two interior microphones were used as stereo input; for exterior detection, the two side-mounted microphones were combined. To adapt the model for three-class classification, the pre-trained attention layers were frozen, and a dense layer with 768 neurons and PReLU activation was added, followed by a softmax output layer with three units [4].

The CRNN model [11,12] combines two-dimensional convolutional layers for spectral feature extraction with a bidirectional GRU layer to model the temporal evolution of sound events. Its frame-wise output allows for temporal localization of events within each segment. In contrast, the YOLO-based model [8] follows a fully convolutional approach adapted from object detection in computer vision. It produces predictions at multiple temporal resolutions using bounding-box-like structures, without relying on sequential modeling. Both CRNN and YOLO operate on single-channel audio, and were thus trained and tested on individual microphone signals. Input Mel spectrograms for these models were computed with 128 Mel bands, a 32 ms window, a 16 ms hop, a Hamming window function, and a 2048-point FFT, all sampled at 16 kHz.

The models were trained on a curated subset of open-

source datasets, primarily based on AudioSet [15], augmented with additional publicly available recordings of sirens, horns, and urban noise. Training procedures were customized for each architecture, ensuring that learned features aligned with the goal of robust acoustic event detection under real-world conditions.

While AST was trained using 4-second monophonic labels, both CRNN and YOLO were originally designed for polyphonic event detection across 10-second clips. For evaluation, however, all models were tested using 4-second segments to ensure consistency. Although AST does not support polyphony, this limitation had minimal impact in our experimental setting, where overlapping events were rare. Nevertheless, this distinction remains an important consideration when interpreting the models' capabilities in more complex acoustic environments.

2.3 Performance Metrics and Evaluation

The performance of each model was evaluated using standard classification metrics: precision (P), recall (R), and F1-score (F1), which are commonly used in classification tasks. These metrics provide a comprehensive assessment of model performance, particularly in scenarios with imbalanced class distributions. In such cases, it is essential to evaluate not only the quantity of correct predictions but also their relevance and consistency. This is especially important in the context of acoustic event detection, where the goal is to identify specific sounds (e.g., horns and sirens) amidst varying levels of background noise.

The recordings used in this study were segmented into 4-second clips with a 1-second sliding window. Each segment was assigned a single label—*Horn*, *Siren*, or *Nothing*—with events outside these categories discarded, particularly for the CRNN and YOLO models. These models generate hard predictions, meaning they provide precise timestamps for detected events within each 4-second segment. For consistency across all models, these timestamps were converted into weak labels, simplifying the task to a binary classification problem, i.e., determining whether an event was present or absent in each segment, without considering its exact duration or temporal position.

Each model was evaluated using both interior and exterior microphones, under two distinct recording conditions. Scenario 1, with minimal background noise, represents a controlled environment, allowing us to assess model performance in an ideal setting. In contrast, Scenario 2 introduces background noise from conversations, radio playback, and mobile phone use, simulating a more realistic in-vehicle environment. By analyzing the performance of the models in these scenarios, we can assess how the placement of the microphone affects the detection accuracy and robustness of the models under varying acoustic conditions.

3. RESULTS

The results presented in Table 1 highlight the impact of microphone placement and background noise on the performance of the three deep learning models—AST, YOLO, and CRNN—in detecting acoustic events. Overall, all models performed better with external microphones compared to internal ones. This effect was observed in both scenarios, but it was especially pronounced in Scenario 2, where background noise from conversations, radio playback, and mobile phone use inside the vehicle further degraded performance.

In Scenario 1, where external microphones captured clearer signals, AST achieved the highest F1-score (0.77), followed closely by YOLO (0.72) and CRNN (0.51). AST's superior performance in this controlled environment is largely attributed to its high recall values, particularly for horn detection (0.96). However, when evaluated inside the vehicle, all models experienced performance drops. AST, despite maintaining the highest F1-score (0.69), showed a notable decline in recall (0.63), while YOLO and CRNN also decreased to 0.50 and 0.34, respectively.

The performance degradation is even more evident in Scenario 2, where additional noise sources such as conversations, radio, and mobile phone sounds were present. AST exhibited the most pronounced decline, particularly for horn detection inside the vehicle, where its precision, recall, and F1-score dropped to zero. This suggests that AST, which assigns a single label per segment, struggles in polyphonic scenarios where multiple sound events overlap, leading to a total drop in performance for horn detection when other, more dominant sources are present. In contrast, while YOLO and CRNN also showed a considerable drop in performance, they retained some detection capability for horns, with F1-scores of 0.15 and 0.17, respectively. However, AST maintained a similar average F1-score inside the vehicle due to its stronger performance in siren detection, despite its inability to detect horns in this scenario.

Although the results of the experiment suggest that external microphone placement might be more effective for detecting sirens and horns, it is important to note that

		Transformer				YOLO				CRNN			
		Scenario 1		Scenario 2		Scenario 1		Scenario 2		Scenario 1		Scenario 2	
		Outside	Inside	Outside	Inside	Outside	Inside	Outside	Inside	Outside	Inside	Outside	Inside
Horn	P	0.58	**0.80**	**0.65**	0.00	**0.82**	0.36	**0.37**	0.16	**0.28**	0.14	**0.21**	0.10
	R	**0.96**	0.87	**0.38**	0.00	**0.78**	0.70	**0.45**	0.17	**1.00**	0.96	**0.76**	0.28
	F1	0.72	**0.83**	**0.48**	0.00	**0.80**	0.47	**0.41**	0.17	**0.44**	0.24	**0.33**	0.14
Siren	P	0.89	**1.00**	**1.00**	**1.00**	0.65	**0.76**	**0.97**	0.61	**0.47**	0.37	0.46	**0.76**
	R	**0.76**	0.38	**0.54**	0.18	**0.63**	0.41	**0.40**	0.08	**0.76**	0.52	**0.51**	0.11
	F1	**0.82**	0.55	**0.70**	0.30	**0.64**	0.54	**0.57**	0.14	**0.58**	0.44	**0.49**	0.19
Avg.	P	0.73	**0.90**	**0.82**	0.50	**0.73**	0.56	**0.67**	0.38	**0.38**	0.26	0.33	**0.43**
	R	**0.86**	0.63	**0.46**	0.09	**0.71**	0.56	**0.43**	0.13	**0.88**	0.74	**0.64**	0.19
	F1	**0.77**	0.69	**0.59**	0.15	**0.72**	0.50	**0.49**	0.15	**0.51**	0.34	**0.41**	0.17

Table 1: Performance metrics (Precision, Recall, and F1-score) for horn and siren detection using AST, YOLO, and CRNN models under two recording scenarios. Results are presented for both external and internal microphones.

this study has excluded certain variables that could influence monitoring effectiveness in real-world conditions. These variables include background noise from the external environment, weather conditions, microphone wear, among others. Additionally, for comprehensive monitoring of the vehicular environment, it is crucial to consider that monitoring the interior may be more complex from the outside due to the vehicle's inherent acoustic insulation, compared to detecting external events from the inside. This suggests that future research could explore the feasibility of implementing two microphone sources, each focused on detecting specific events from the interior and exterior, which could improve the accuracy and reliability of the system in more complex scenarios.

4. CONCLUSIONS

This study evaluated the performance of three deep learning models—AST, YOLO, and CRNN—for detecting external acoustic events (such as sirens and horns) in vehicular environments under various conditions. The results highlight the significant impact of microphone placement and background noise on detection accuracy. External microphones generally provided clearer signals, leading to better performance, while internal microphones were more susceptible to degradation, particularly in noisy conditions.

AST achieved the highest F1-score in controlled environments but struggled with overlapping sounds, resulting in a complete failure in horn detection inside the vehicle. In contrast, YOLO and CRNN maintained some detection ability in these challenging scenarios, although with lower overall accuracy.

Future work should explore the models' performance in more complex tasks, such as detecting the precise onset and duration of events, rather than merely classifying their presence within fixed segments. Additionally, investigating multi-microphone fusion strategies could help mitigate the limitations observed with single-channel inputs.

5. ACKNOWLEDGMENTS

This work has been supported by Grant TED2021-131003B-C21 funded by MCIN/AEI/10.13039/501100011033 and by the "EU Union NextGenerationEU/PRTR", as well as by Grant PID2022-137048OB-C41 funded by MICIU/AEI/10.13039/501100011033 and "ERDF A way of making Europe". Authors would like also to thank *Generalitat Valenciana-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).

6. REFERENCES

[1] K. Choudhury and D. Nandi, "Review of emergency vehicle detection techniques by acoustic signals," *Transactions of the Indian National Academy of Engineering*, vol. 8, 2023.

[2] L. Marchegiani and P. Newman, "Listening for sirens: Locating and classifying acoustic alarms in city scenes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, 2022.

[3] M. Y. Shams, T. A. El-Hafeez, and E. Hassan, "Acoustic data detection in large-scale emergency vehicle sirens and road noise dataset," *Expert Systems with Applications*, vol. 249, 2024.

[4] L. Banchero, F. Vacalebri-Lloret, J. M. Mossi, and J. J. Lopez, "Enhancing road safety with AI-powered system for effective detection and localization of emergency vehicles by sound," *Sensors*, vol. 25, no. 3, 2025.

[5] S. Grimm and J. Freudenberger, "Wind noise reduction for a closely spaced microphone array in a car environment," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2018, 2018.

[6] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2021-June, 2021.

[7] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, 2021.

[8] C. Castorena, M. Cobos, J. Lopez-Ballester, and F. J. Ferri, "A safety-oriented framework for sound event detection in driving scenarios," *Applied Acoustics*, vol. 215, p. 109719, 2024.

[9] M. Zohaib, M. Asim, and M. ELAffendi, "Enhancing emergency vehicle detection: A deep learning approach with multimodal fusion," *Mathematics*, vol. 12, no. 10, 2024.

[10] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-Supervised Audio Spectrogram Transformer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 10699–10709, Jun. 2022.

[11] F. Ronchini, R. Serizel, N. Turpault, and S. Cornell, "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)*, (Barcelona, Spain), pp. 115–119, 2021.

[12] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, p. 0, 2019.

[13] C. Castorena, J. Lopez-Ballester, J. A. D. Rus, M. Cobos, J. Lopez-Ballester, and F. J. Ferri, "Edge Computing for Driving Safety: Evaluating Deep Learning Models for Cost-Effective Sound Event Detection," *J Supercomput*, vol. 288, 2025.

[14] Y. Gong, Y. A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 1, 2021.

[15] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.