Interspeech 2024
1-5 September 2024, Kos, Greece

# *Sound of Traffic*: A Dataset for Acoustic Traffic Identification and Counting

*Shabnam Ghaffarzadegan, Luca Bondi, Wei-Chang Lin, Abinaya Kumar, Ho-Hsiang Wu, Hans-Georg Horst, Samarjit Das*

Bosch Center for Artificial Intelligence

first.last@bosch.com

## Abstract

We introduce *Sound of Traffic*, the largest publicly available dataset for traffic identification and counting to date. With over 415 hours of multichannel acoustic traffic data recorded in six different locations, it encompasses varying levels of traffic density and environmental conditions. In this work, we discuss strategies for automatic collection and alignment of large amount of labeled data, leveraging existing asynchronous urban sensors such as radar, cameras, and inductive coils. In addition to the dataset, we propose a simple baseline system for vehicle counting divided by type of the vehicle (passenger vs. commercial vehicle) and direction of travel (right-to-left and left-to-right), a fundamental task for traffic analysis. The dataset and baseline system serve as a starting point for researchers to develop more advanced algorithms and models in this field. The dataset can be accessed at https://zenodo.org/records/10700792 and https://zenodo.org/records/11209838.

**Index Terms**: Sound of Traffic dataset, Traffic counting and identification, Automatic labeling, Audio event detection

## 1. Introduction

Automatic urban sound understanding is the task of classifying sounds in urban areas, their location and direction of movement. Urban audio analytics has recently gained lots of attention in both academia and industry with great potentials for smart city applications such as traffic monitoring and urban noise mapping [1–5]. These systems are designed based on several different sensors grouped into two main categories of intrusive and non-intrusive sensors. Examples of intrusive sensors that are embedded in the road are induction loops, vibration or magnetic sensors. Examples of non-intrusive systems mounted over or on the side of the road are radar, cameras, infrared, acoustic sensors, and GPS-based systems. Acoustic sensors offer numerous advantages over other sensor options, making them a highly desirable choice either on their own or in combination with other sensors. These advantages include easy installation (can be installed on poles and traffic light), maintenance, affordability, power efficiency, wide coverage, and resilience in adverse weather and low-visibility conditions.

One of the main challenges in developing acoustic-based traffic monitoring systems is having access to temporally labeled audio data in scale. Publicly available datasets specifically focused on acoustic-based traffic monitoring (i.e. counting, vehicle type classification, direction of travel estimation) are relatively scarce. A few datasets such as TUT Urban Acoustic Scenes 2017 and 2018 [6, 7], SONYC [8], FSK50k [9] and AudioSet [10] provide general traffic-related sound classes such as car, truck, train, metro, tram, etc. Although these datasets lack specific annotations for parameters like counting, direction of travel, or vehicle type, they can still serve as a foundation for developing algorithms and models in these areas.

In addition to the previously mentioned datasets, there are three publicly available datasets that provide more extensive annotations in the field of acoustic-based traffic monitoring. MIVIA road [11] contain recordings of urban environments, capturing various sounds like engine noise, crashes, and tire skidding. MAVD-Traffic dataset [12] provides detailed annotations for categorizing vehicle types, such as cars and buses, as well as vehicle components like engines and brakes. Additionally, it includes annotations for various vehicle actions, including idling and accelerating. Finally, IDMT-Traffic [13] is one of the most comprehensive resources available for acoustic traffic monitoring research. It features 15,706 of 2-second stereo audio clips recorded in four different locations. The dataset is annotated for vehicle type and direction of travel using a camera setup right next to the microphones.

In this paper, we introduce *Sound of Traffic*, the largest publicly available dataset for acoustic traffic monitoring task. The data consists of 415 hours of labeled audio segments collected over several years in various traffic conditions, acquired via a 4-channel linear microphone array. A portion of the dataset is automatically annotated utilizing existing sensors such as radar, camera and inductive coil, enabling large-scale annotation. The contributions of this work are three folds:

1. *Data collection:* We introduce *Sound of Traffic* in section 2, a unique acoustic dataset designed to promote and support research in acoustic-based traffic monitoring solutions. *Sound of Traffic* consists of 24,900 1-minute labeled audio segments collected via 4-channel linear microphone array in various traffic conditions from light to heavy traffic.
2. *Data annotation*: Manual annotation of acoustic traffic sounds is a challenging task for humans, as accurately labeling factors such as car type, vehicle count, and speed solely through auditory perception is exceedingly difficult to impossible. To address this challenge and enable annotation on a larger scale, we leverage the capabilities of existing sensors already deployed in urban environments to automatically annotate audio data for training purposes. In section 3, we propose a new automatic data annotation technique that enables the seamless alignment of audio recordings with labels derived from a diverse range of sensors including radar, cameras, and inductive coils.
3. *Baseline system*: Finally, in section 4 we present the results of benchmark experiments for vehicle counting divided by type of the vehicles and their direction of travel using a simple convolutional neural network trained without any external data.



10.21437/Interspeech.2024-1205

Figure 1: *A few examples of microphone array locations in urban environment. The exact sensor location is indicated by a red circle.*

## 2. *Sound of Traffic* Dataset Overview

**Recording procedure** To collect traffic sound data, a linear microphone array consisting of four MEMS microphones with a 24cm aperture is installed on the roadside, positioned parallel to the direction of travel. The installation of the microphone array varies depending on the available environment, such as poles and traffic lights. This results in arrays being installed at different heights and distances from the street side for each location. We acquired audio samples from six distinct locations across Europe and the United States. Figure 1 shows some installations. These sites were carefully chosen to include a diverse range of traffic conditions and environments. This collection includes both rural and urban areas, covering a spectrum of traffic densities ranging from country roads with a maximum of 5 vehicles per minute to intercity roads with up to 30 vehicles per direction per minute, covering the period from November 2019 to December 2022. A voice activity detector is always monitoring the audio stream, and recording is paused if any voice is detected. The continuous audio stream is chunked into segments of duration of 60 seconds.

To ensure reproducibility and facilitate further research, we released a subset of the collected data in pre-defined train and validation splits at https://zenodo.org/records/10700792. The test split data can be found at https://zenodo.org/records/11209838. Figure 2 provides an overview of the amount of data available for each split for the six locations. For locations 1, 3, and 6 automatic labeling was done with utilizing cameras, coil, or radar sensors, making a sizeable amount of labeled data available. Data from location 2, 4, and 5 was manually labeled while standing on the side of the road, thus the limited availability of data from these locations. Additionally, to gain a comprehensive understanding of the data statistics, we present a summary of the data distribution for each location and various hours of the day. Figure 3 shows the distribution of vehicle counting per label per location. Furthermore, Figure 4 illustrates the traffic distribution for three specific locations, categorized by the hour of the day.

**Reference labels** We define the traffic monitoring task with three specific objectives. The first is to count the number of vehicles passing by within a one-minute segment. The second is to identify the direction of travel. The third is to classify the type of vehicle, specifically distinguishing between passenger cars (referred to as "car") and commercial vehicles (referred to as "cv").

To facilitate this task, we provide ground truth labels that encompass counting vehicles in four distinct classes within a 60 second segment. In the following, when we refer to *left* and *right* we assume we are looking at the road from behind the

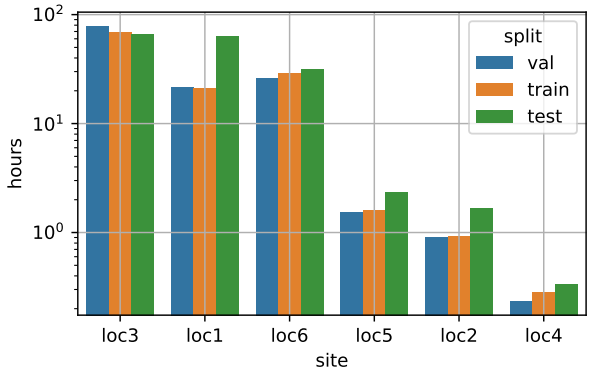


Figure 2: *Amount of data in hours per each site and split. The y-axis is on a logarithmic scale, to accommodate for the very different amount of data available across sites and splits, ranging from less than an hour to almost 100 hours.*

microphone array. These labels are as follows: *car/left*: number of cars heading left; *car/right*: number of cars heading right; *cv/left*: number of commercial vehicles heading left; *cv/right*: number of commercial vehicles heading right.

**Meta-data** In addition to audio recordings, we also provide various meta-data, including: 1) *sensor location ID*: this allows for the development of different models for each site, accommodating specific characteristics and conditions; 2) *Timestamp*: we provide the date and time information, enabling models to learn and adapt to varying traffic conditions that naturally occur at different times, such as weekdays or weekends, rush hours, and other time-specific patterns; 3) *Array position information*: we offer details on the position of the sensor relative to the traffic lanes such as height and distance to the road, aiding in understanding the context of the data; 4) *Maximum speed of vehicles*: we provide information on the highest speed recorded at the specific location; 5) *Highest number of pass-by vehicles*: we provide data regarding the highest number of recorded vehicles pass-by in each direction for each location.

## 3. Automatic data annotation

One of the main challenges in developing traffic monitoring systems is having access to temporally labeled audio data, as most existing datasets have limited or weakly labeled events, i.e., labels without the temporal information. Temporal tagging of audio events is a very expensive task which requires huge human effort and is prone to errors.

In this work we leverage auxiliary sensors already installed in urban settings to automatically label audio data. Example of auxiliary sensors are cameras, induction coils, and radars. Auxiliary sensors generate a stream of events for each pass-by, identifying the vehicle type, direction, and optionally speed.

One source of misalignment between the acoustic fingerprint of a pass-by and the events collected by an auxiliary sensor is the relative position between the microphone array and the other sensor. For example, an induction coil installed on the road a few meters away from the microphone array will record pass-bys in one direction with some delay, while pass-bys in the other direction ahead of time. Geometry-based misalignment are typically easy to compensate for, given the vehicle speed, direction of travel, and information about the geometry of the

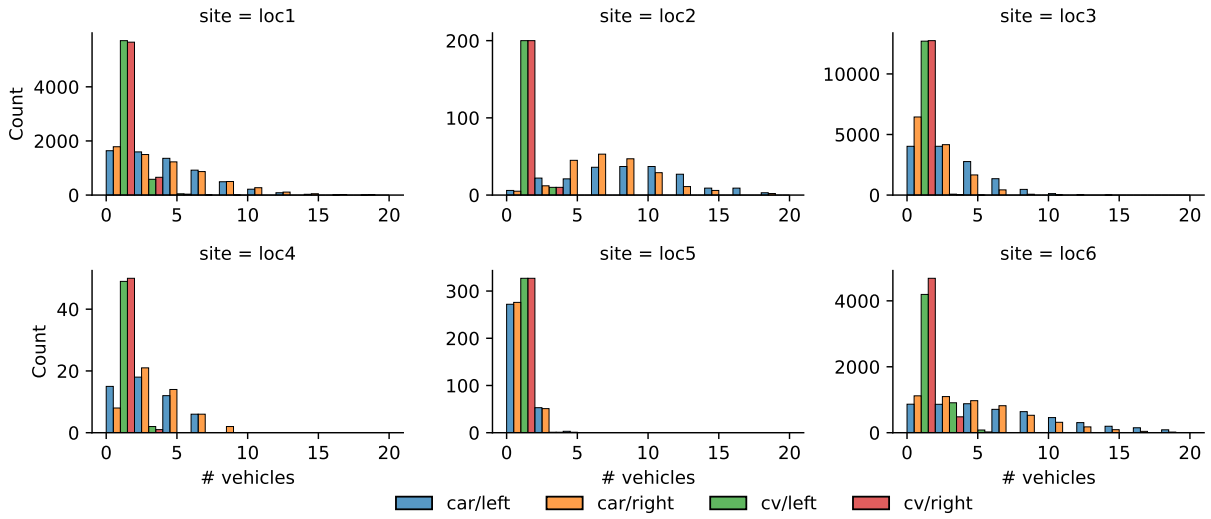


Figure 3: *Number of one minute segments (Count) vs. number of vehicles passing by (# vehicles), shown per site and label. Locations 1, 4, and 6 are extra-urban roads. Locations 2 and 3 are a urban roads, close to an intersection. Location 5 is a private road in a corporate campus.*

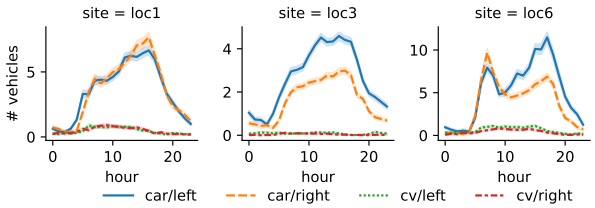


Figure 4: *Traffic distribution per label over 24 hours, for three locations that have at least 24 hours of data.*

site. All auxiliary sensors used in this work provide these information.

The second source of misalignment and the main challenge of using auxiliary sensors is the lack of clock synchronization between the sensor and the audio stream. This is mainly due to the different clock domains of isolated devices, hence an offset and a drift between the audio stream and the auxiliary sensor stream need to be taken into account. Assuming that one time per day both the microphone array and the auxiliary sensor are synchronized to a common Network Time Protocol server, in the following we describe a correlation-based method used in this work to compensate for offset and drift.

**Automatic drift and offset compensation**

1. The system takes as input a list of events $E = \{\hat{e}_i\}$ generated by an auxiliary sensor (induction coil, radar, camera) and an audio segment (wave form), see Fig. 5 (top).
2. A simple vehicle detector, based on Mel Frequency Cepstrum Coefficient (MFCC) followed by Logistic Regression (LR), computes the likelihood $L(t)$ of a vehicle passing by, over time, see Fig. 5 (bottom). The LR is trained on just a few samples for each location.
3. An indicator function $I(t)$ is generated from the stream of events, by applying a Gaussian point spread function to each event in $E$, see Fig. 5 (middle).
4. The shift between the audio stream and the events stream is computed via cross-correlation between the likelihood of a vehicle passing by ($L(t)$) and the indicator function ($I(t)$) generated from the auxiliary sensor events. The shift-compensated events $\tilde{e}_i$ are shown in red in Fig. 5 (top).
5. Shifts computed for different segments at different time of the day are fed to a RANSAC regressor, to estimate the average drift and offset between the auxiliary sensor and the audio stream for the day under analysis.

## 4. Traffic Monitoring Baseline

The proposed baseline system takes as input 4 channels of raw audio and computes 6 channels of Generalized Cross-Correlation with Phase transform (GCC-Phat) features, as well as 4 channels of Log Mel Spectrogram. Fig. 6 shows an example of the two representations, with GCC-Phat carrying significant information about the direction of travel. GCC-Phat and Log Mel Spectrogram are independently processed by batch normalization and two convolutional blocks with 16 and 32 filters, kernel size 3, stride 2, then averaged across the channels dimension, and further processed by two time-distributed linear blocks with 32 and 64 filters. The two branches are then concatenated along the feature dimension, and processed by three time-distributed linear blocks with 64, 128, 256 filters. The summation over time goes through batch normalization and is finally fed to a linear layer with 4 output neurons and ReLU activation, providing the count of vehicles for the four classes *car/left*, *car/right*, *cv/left*, *cv/right*. The output of the model is not restricted to integer numbers, thus decimal output values are possible to indicate uncertainty on the prediction. The model has a total of 69.4K trainable parameters.

**Metrics** In the context of a regression problem, we employ both distance-based and correlation-based metrics to assess the performance of the system from different perspectives. The distance-based metric directly calculates the regression errors, while the correlation-based metric accounts for the scaling fac-

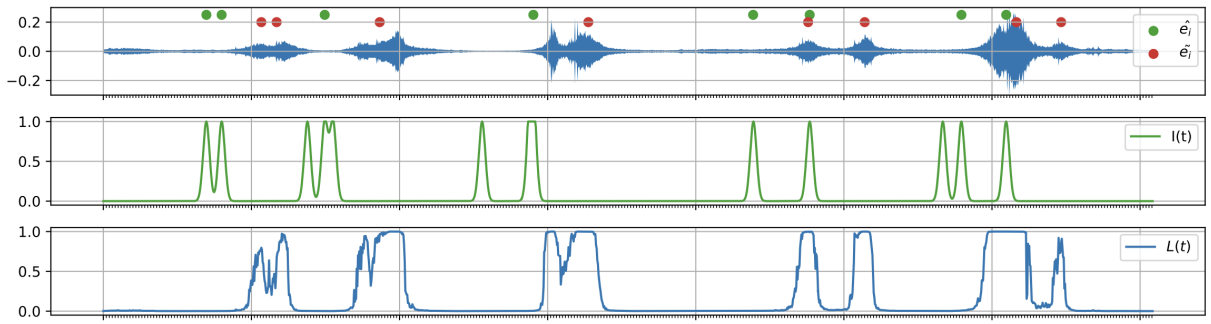


Figure 5: ***top**: Waveform of an audio segment overlaid with events from the auxiliary sensor ($\hat{e}_i$ / green), and events after shift correction ($\tilde{e}_i$ / red). **middle**: Indicator function ($I(t)$) obtained applying a Gaussian window function to the geometry-compensated labels. **bottom**: Likelihood function ($L(t)$) of a vehicle being present in the audio stream.*

| | *car/left* | | *car/right* | | *cv/left* | | *cv/right* | |
|---|---|---|---|---|---|---|---|---|
| | KTau | RMSE | KTau | RMSE | KTau | RMSE | KTau | RMSE |
| **loc1** | 0.49 | 2.49 | 0.49 | 2.83 | 0.20 | 0.77 | 0.26 | 0.70 |
| **loc2** | 0.42 | 4.31 | 0.48 | 2.32 | 0.16 | 0.46 | -0.15 | 0.61 |
| **loc3** | 0.56 | 1.79 | 0.55 | 1.25 | 0.16 | 0.29 | 0.20 | 0.20 |
| **loc4** | 0.46 | 1.93 | 0.66 | 1.36 | 0.27 | 0.51 | 0.60 | 0.51 |
| **loc5** | 0.07 | 1.03 | 0.20 | 0.94 | 0.15 | 0.23 | 0.18 | 0.21 |
| **loc6** | 0.83 | 1.76 | 0.73 | 1.79 | 0.69 | 0.61 | 0.60 | 0.50 |

Table 1: *Baseline results for the six sites. A lower Root Mean Squared Error (RMSE) is better. A higher Kendall's Tau Coefficient (KTau) is better.*

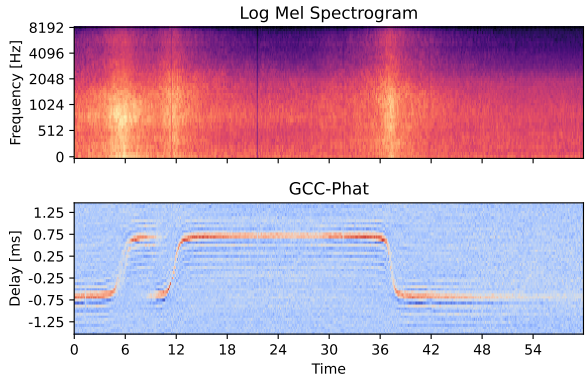


Figure 6: *Baseline system input representation example. GCC-Phat is computed between the two external microphones.*

tor and considers the overall upward or downward trend. A well-performing system is expected to capture both the local variations and the global trends in traffic counting. To evaluate the system, we utilize the following metrics:

- **Root Mean Square Error (RMSE)**: a commonly used metric in regression tasks, which quantifies the prediction errors using the Euclidean distance. RMSE ranges from 0 to infinity. Unlike Mean Absolute Error (MAE), RMSE assigns more weight to larger errors, providing a more sensitive measure of the overall prediction accuracy.
- **Kendall's Tau Rank Correlation (KTau)**: measures the ordinal association between two quantities. It is often preferred over other metrics like Pearson's or Spearman's correlation due to its robustness against outliers. Kendall's Tau can range from -1 to 1, indicating the strength and direction of the association. A value closer to one indicates that the system is predicting the correct trend.

**Results** The baseline system is trained using data from all the locations. Training data comes from the train splits, validation data from the validation splits. Optimization is driven by the Adam optimizer, with a learning rate of 0.002, a batch size of 128 segments, on a single NVIDIA Tesla V100 with 32GB of VRAM. The model with the smallest validation loss is selected for inference. Evaluation is performed on the test split. Results in Table 1 show the importance of the two complementary metrics. *loc5* has on average the smallest RMSE, however KTau is also very small, indicating that the system is not responsive to traffic variations in the site. This happens as the site has a very low traffic rate, and the system mostly predicts values between 0 and 1. The RMSE is small due to the easy traffic conditions, however a low KTau shows that the system does not respond reliably to higher traffic with higher counting. On the other end, *loc6* has a high RMSE as well as the highest KTau. The site has the highest traffic density (see Fig. 3) and the system tracks quite well the variations in traffic (high KTau). However, due to the moderate traffic conditions, the RMSE for the site is also high, signaling that the punctual estimation could be improved.

## 5. Conclusion

In this paper we introduced *Sound of Traffic*, a new dataset for acoustic traffic counting. Featuring around 415 hours of real-world data collected from 6 sites with a 4-channel linear microphone array. This dataset enables research on traffic identification and counting based on microphone array signals in realistic scenarios. Ground truth vehicle counts per vehicle type and direction of travel are provided for train, validation, and test splits, to allow autonomous research and evaluation. A baseline system and two metrics are proposed to provide a reference for future works.

# 6. References


[1] M. Ashhad, U. Goenka, A. Jagetia, P. Akhtari, S. K. Ambat, and M. Samuel, "Improved vehicle sub-type classification for acoustic traffic monitoring," 2023.

[2] Y. Na, Y. Guo, Q. Fu, and Y. Yan, "An acoustic traffic monitoring system: Design and implementation," *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, pp. 119–126, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:16168573

[3] Z. Ye, W. Wang, X. Wang, F. Yang, F. Peng, K. Yan, H. Kou, and A. Yuan, "Traffic flow and vehicle speed monitoring with the object detection method from the roadside distributed acoustic sensing array," *Frontiers in Earth Science*, vol. 10, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/feart.2022.992571

[4] O. Ghaffarpasand, A. Almojarkesh, S. Morris, E. Stephens, A. Chalabi, U. Almojarkesh, Z. Almojarkesh, and F. D. Pope, "Traffic noise assessment using intelligent acoustic sensors (traffic ear) and vehicle telematics data," *Sensors*, vol. 23, no. 15, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/15/6964

[5] K. Marciniuk and B. Kostek, "Machine learning applied to acoustic-based road traffic monitoring," *Procedia Computer Science*, vol. 207, pp. 1087–1095, 2022, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050922010468

[6] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 85–92.

[7] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840

[8] M. Cartwright, J. Cramer, A. E. M. Mendez, Y. Wang, H.-H. Wu, V. Lostanlen, M. Fuentes, G. Dove, C. Mydlarz, J. Salamon, O. Nov, and J. P. Bello, "Sonyc-ust-v2: An urban sound tagging dataset with spatiotemporal context," 2020.

[9] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: An open dataset of human-labeled sound events," 2022.

[10] 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[11] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 279–288, 2016.

[12] P. Zinemanas, P. Cancela, and M. Rocamora, "Mavd: A dataset for sound event detection in urban environments," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204854206

[13] J. Abeßer, S. Gourishetti, A. Kátai, T. Clauß, P. Sharma, and J. Liebetrau, "Idmt-traffic: An open benchmark dataset for acoustic traffic monitoring research," 2021.