DOI: 10.61782/fa.2023.0858

forum acusticum 2023

MODELS FOR VEHICLE DETECTION FROM NOISE MEASUREMENTS IN SPARSE ROAD TRAFFIC

Siddharth Venkataraman[1*] **Romain Rumpler**[1]
Elias Ekberg[2] **Kevin Golshani**[2]
[1] The Centre for ECO[2] Vehicle Design and Digital Futures,
The Marcus Wallenberg Laboratory for Sound and Vibration Research (MWL),
KTH Royal Institute of Technology
[2] Department of Mathematics,
KTH Royal Institute of Technology

ABSTRACT

Road traffic noise calculations require modeling the traffic flow in a road network. The reliability of these calculations can be improved with accurate estimation of the traffic flow, including estimation of its temporal variations. Low-cost noise sensors that run on single-board computers in a noise monitoring network are suitable candidates to simultaneously estimate the local temporal traffic flow from their pass-by measurements, using an on-board traffic flow estimator model. Aside from this model requiring to be computationally efficient, it should also be robust, e.g., invariant to sensor position relative to the source, weather conditions, etc. With noise measurements as an input, different noise features and prediction models are tested for vehicle detection. The accuracy of these models is evaluated using traffic count data obtained from dedicated vehicle-counting infrastructure at the locations of the noise sensors. The analysis is restricted to sparse traffic conditions in this initial study.

Keywords: *sound event detection, road traffic, noise measurements*

**Corresponding author*: *sidven@kth.se.*

1. INTRODUCTION

Urban noise pollution may be assessed through direct noise measurements, or noise propagation simulations [1]. Noise measurements provide accurate assessment of the noise levels, but are limited in their spatial resolution. On the other hand, noise simulations enable assessment over wider areas, but their accuracy is dependent on the propagation models and their input data.

The reliable estimation of the traffic flow rate directly influences the reliability of the noise propagation output, since the noise source strength is proportional to the flow rates. An accurate estimation of traffic flow is of particular importance during sparse traffic and low background noise, because the transient nature of noise from individual vehicles is more significant in sparse traffic. It is therefore of value to obtain reliable estimates of traffic flow rates, with high temporal resolution during sparse traffic.

A methodology to improve the assessment of traffic flow rates, and the subsequent noise simulations, is through the use of low-cost noise sensors as sources of local traffic counts. These traffic flow parameters can then be used as inputs to traffic simulation models, which in turn support noise simulation models. The same sensors can also provide noise level measurements, which is their primary purpose, that can be used to validate the output of the simulations. Doing so makes the sensor useful in both the upstream (input to traffic-simulation models) and downstream (validation of noise-simulation models) of the noise assessment methodology.

This paper seeks to present vehicle detection models that use data from the noise sensors for estimation of traf-

fic flow parameters. The noise sensors are part of a local noise sensor network in Stockholm, Sweden. The scope is here initially restricted to sparse traffic conditions, i.e., there is no overlap in the pass-by of adjacent vehicles. Furthermore, the noise data is complemented by weak labels in the form of approximate timestamps of vehicle pass-by detection, obtained from a traffic sensor adjacent to the noise sensor.

2. CASE STUDY

A section of a highly-trafficked road in Central Stockholm, Hornsgatan, was chosen as the test-bed for development of the vehicle detection models. Noise sensors were installed at four locations that were positioned either directly beside the road section, or close to it. These locations are shown in Fig. 1. Two of these four locations also included a traffic sensor. The exact locations of the sensors were constrained by the availability of power and internet infrastructure.

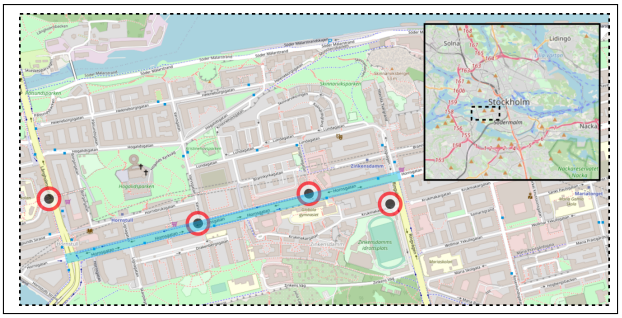

Figure 1. Test-bed along Hornsgatan highlighted in blue, along with location of four noise sensors marked by concentric circles. The two sensors on the right side of the map also included a traffic sensor. Background map obtained from OpenStreenMap [2].

2.1 Sensor infrastructure

Each noise sensor was equipped with a miniDSP Umik-1 [3], a 6 mm electret omni-directional USB measurement microphone. The noise data was measured and processed by a Raspberry Pi Model 3B single-board computer.

Each traffic sensor was equipped with a designated radar sensor for each lane at that location. The sensors for the two lane directions were installed on opposite sites of the road. Each sensor provided a timestamp when a vehicle passes a particular section of the lane.

The installation of the traffic and noise sensors was carried out under the constraints of available mounting points. Therefore, the particular section of the lane that each radar sensor monitored was neither side-by-side to each other, nor to the section of the road closest to the noise sensor. Therefore, the timestamps from the radar sensors were at a lane-specific offset to the noise peak of the vehicle pass-by. Therefore, the radar sensors provided an approximate timestamp of the vehicle pass-by with respect to the noise sensor data at that location.

2.2 Sparse traffic

The focus of this paper was restricted to sparse traffic conditions. Sparse traffic in this paper was defined as traffic conditions where the distance between two successive vehicles on a road was large enough that there would be no or negligible overlap in their impact on the road and noise sensor data.

The availability of road sensor data enabled identifying time periods where such sparse conditions exist. In this paper, a minimum duration of 20 seconds between two radar-detected vehicles was required for the prevailing traffic conditions to be considered sparse.

3. METHOD

A vehicle detection model is developed for identifying vehicle pass-by from noise measurements during sparse traffic conditions, taken at a noise sensor adjacent to a road. The model is trained to predict strong labels for a vehicle pass-by, i.e., timestamps denoting the start and stop of the pass-by event. Along with the noise measurement, the training data also includes weak labels, i.e., approximate timestamps of a vehicle pass-by, obtained from a traffic sensor.

The development of the vehicle detection model is composed of two parts:

1. Generation of strong labels
2. Prediction of strong labels

The noise measurements from the sensors are converted into spectrograms. These spectrograms are processed further to obtain other noise features; see Section 3.1. An unsupervised classification technique is used for converting the weak labels into strong labels; see Section 3.2. Lastly, the noise features along with the generated strong labels are used as training data for a supervised classification technique, described in Section 3.3. This

classifier is then used to predict the onset and offset of a vehicle pass-by from an input noise measurement, and the output is presented in Section 4.

3.1 Feature extraction

The time-domain noise measurements from the noise sensor, taken at a sampling frequency of 22050 Hz, are converted to frequency-domain absolute-magnitude spectrograms with a frequency resolution of about 5.4 Hz.

The spectrogram, and multiple features extracted from it are tested as possible input candidates for the vehicle-detection model. In the presented assessment, the log-mel spectrogram and the Mel-frequency cepstrum coefficients (MFCCs) are evaluated and used as features to represent the noise measurements. The noise data is converted to log-mel spectrograms with 64 bands, and MFCCs up to the 13^{th} coefficient. These features are chosen because of their known ability to effectively differentiate and recognize audio based on physiological characteristics of the human ear [4]. Although other features may be more relevant for vehicle generated noise, the model presented in this paper uses these features as a preliminary input to the detection model, making it possible for future work to implement and test more advanced features. For the sake of clarity, only the results from using the 13 MFCCs are presented.

3.2 Strong label generation

The weak labels from the traffic radar sensor are converted to strong labels using the k-means classification technique. The weak labels, being approximate timestamps of a vehicle pass-by, are used as a mid-point for extracting a relevant section of the noise features from the entire noise measurement dataset. An offset of 15 seconds about this mid-point (i.e., a 30-second duration window) is defined, in order to include data corresponding to when noise from the vehicle pass-by is clearly absent. Note that only sparse vehicle events are chosen as input for this model, i.e., in each window there is only one vehicle present.

Each time-step in the window is then classified using a k-means classifier into one of 3 possible clusters, with classification based on its features, i.e., the 13 MFCCs calculated for each time-step. The choice to cluster using three clusters is chosen after a process of trial-and-error. Using only 2 clusters is found to be insufficient to capture the vehicle pass-by, likely because noise from other sources prevents the clustering algorithm from clustering the pass-by time events under a single cluster. Using more than 3 clusters results in identifying spurious events, therefore adding an additional challenge of identifying the right cluster that corresponds to the real pass-by event.

The resulting clusters are low-pass filtered across the time domain to divide the measurement window into continuous clusters. A low-pass filter is required because the cluster labels rapidly fluctuate around the time of vehicle onset and offset. The low-pass filter removes these oscillations, allowing for a more realistic segmentation of the measurement window.

Finally, the dominant cluster in the middle of the window is identified as the cluster containing the pass-by, and the boundaries of this cluster yield the strong labels for this vehicle pass-by event. Note that the weak label need not necessarily be contained within the strong label, as the purpose of this weak label is only to serve as a mid-point for the noise data window. An example of the resulting strong label for one such window is shown in Fig. 2.

The strong label defines a time window that begins at the vehicle onset and ends at the vehicle offset. Since this label is defined as the boundary of a cluster, a definition of a vehicle pass-by based on changes in sound levels is not explicitly required.

3.3 Event onset and offset prediction

A random forest classifier is used for predicting the strong labels for a given noise measurement. This classifier is trained on the noise features of a measurement window used to generate the strong labels (as described in Section 3.2). The random forest is trained using 500 trees, and uses bootstrapping. The features for each time-step is individually classified by the trained classifier as either inside a pass-by event or outside. The predictions are then processed through a low-pass filter, like in the case of the strong label generator in Section 3.2, to obtain the final predictions of onset and offset.

The accuracy of these predictions is assessed using a 10-fold cross validation between the predictions and the generated labels. Two kinds of methods are applied to calculate these metrics: frame-by-frame and collar-based. The frame-by-frame method compares the predicted and the generated label one time-step at a time. On the other hand, the collar-based method [5] allows for the predicted and the generated strong label of an entire pass-by event to pass only if the predicted onset is within 1 second of the

generated onset, and the predicted offset is within a collar around the generated offset. This collar around the offset is taken as 50% of the generated label's total duration.

Predictions from a trained classifier are shown in Fig. 3 for a single case, and Fig. 4 for all cases combined. A prediction for a pass-by event is obtained using a model trained on 90% of the entire data set, and this training set excludes the measurement that is to be predicted.

4. PRELIMINARY RESULTS

The strong label generator and predictor models are trained on data from sparse traffic, obtained over a 24-hour period from a location in Fig. 1 having both a noise sensor and traffic radar sensor. On this chosen 24-hour period, a total of 20,681 vehicles were counted by the radar sensors on the four lanes. The requirements for sparse traffic conditions, i.e., at least 20 seconds between two consecutive vehicles, reduces the number of sparse events to about 0.8% of the total. Noise measurements for the remaining 143 vehicle pass-by events, obtained using weak labels from the radar sensor, are used for generating the following results.

4.1 Strong label generation

In Fig. 2 is an example of the output from the strong label generator described in Section 3.2.

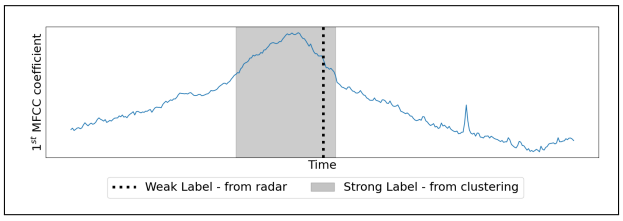

Figure 2. Example of output from the strong label generator. Label generator trained on 13 MFCCs from each time-step of the noise data window.

The noise data window is represented using the first MFCC, which correlates to the total energy in the signal at that time-step. The weak label is the approximate timestamp of vehicle pass-by obtained from the traffic radar sensor. The strong label is the onset and offset timestamp of a vehicle pass-by, represented as a shaded region of the noise data window.

The strong label in Fig. 2 contains the vehicle pass-by peak, represented through the first MFCC, and is about 5.9 seconds in duration (within the 30 seconds window duration).

Strong labels are generated for the 143 pass-by events. Out of these, 5% of the measurements did not produce labels that are realistic, since the pass-by is detected at the edges of the measurement instead of at the center. The remaining 135 strong labels are shown in Fig. 4.

4.2 Strong label prediction

In Fig. 3 is an example of the output from a trained strong label predictor described in Section 3.3. This predictor is trained using a training-set, and an example from the testing-set is presented in Fig. 3. The strong label predictions for all the pass-by events considered are shown in Fig. 3.

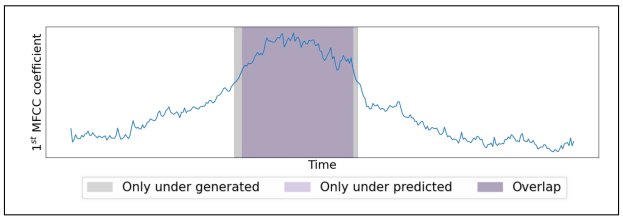

Figure 3. Example of output from the strong label predictor. Predictor trained on MFCCs from the noise data training-set. Expected (in grey) and predicted (in purple) strong labels shown for a representative noise data window

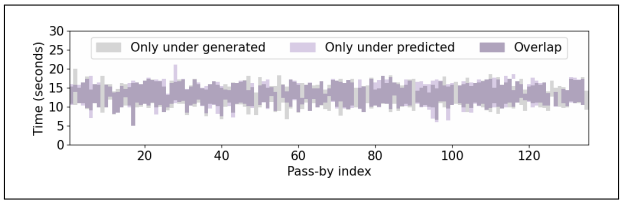

Figure 4. Generated (in grey) and predicted (in purple) strong labels for all sparse traffic pass-by events considered in a 24-hour period. Each column corresponds to a single pass-by event, and the colored bars annotate the generated and the predicted strong labels, with an overlap denoting a measurement window in which a vehicle pass-by is identified by both the generated and the predicted labels

In Fig. 3, similar to Fig. 2, the first MFCC is used

to represent the energy in the noise signal under consideration. Two strong labels are presented in Fig. 3. The predicted strong label, highlighted in purple, is the output of this model, when given as input the MFCCs corresponding to this particular noise data window. The generated strong label, highlighted in grey, is obtained from the strong label generator (Section 3.2), and is used for testing the accuracy of the predicted label. The predicted label is shorter in duration but corresponds well with the generated label, overlapping by about 90% with the latter. The predicted and generated labels correspond to a duration of about 6.7 seconds and 7.4 seconds, respectively, within the 30 seconds window duration.

In Fig. 4, the generated and the predicted labels for the 135 pass-by events are shown. Out of the 135, only 125 labels are shown because the predictor failed to identify a pass-by event for 10 of the cases.

A 10-fold cross validation of the predicted and generated strong labels from the 135 pass-by events yielded an average accuracy in the frame-by-frame method to be 0.93, and in the collar-based method to be 0.62.

5. DISCUSSION

The preliminary results presented in Section 4 show scope for the strong label predictor model to detect vehicles in sparse traffic conditions.

The accuracy of the predicted labels in Fig. 4 depends on the choice of evaluation. A frame-by-frame evaluation gives a high accuracy of 0.93. On the other hand, a collar-based evaluation yields a lower accuracy of 0.62. The evaluation under the collar-based method is more strict, since a prediction contributes to the score only if all the frames-wise predictions satisfy the collar requirement. In comparison, for the frame-by-frame evaluation, even the absence of any predicted event within a window can have a non-zero score through the true-negatives.

The weak labels enable extracting a relevant noise data window, which is then used by the strong label generator to provide input training data for the strong label predictor. The label predictor, once trained, can be used to identify vehicle onset and offset timestamps from unlabeled noise data, i.e., data from noise sensors without a traffic radar sensor coupled to it.

The noise features used for this analysis are restricted to MFCCs. Although these coefficients allow for training the models, other features can be considered, e.g., those identified through feature learning [6], which may more effectively capture the frequency content and temporal dynamics typical for a vehicle pass-by.

The noise data in this analysis is obtained from a single noise sensor and from the same 24 hour period, thereby overfitting the models to variables specific to this temporal and spatial environment. To increase the model robustness, the input data could be expanded to include data from other noise sensors, and data from a longer time period. Another solution could be to include other independent and external sources of noise data [7].

The current requirement for sparse traffic places strict constraints on the practical applicability of these models. To circumvent this, more elaborate label generation and prediction models may be implemented, such as CNNs [8], and student-teacher models [9].

Development of these detection models may increase their robustness and reliability to detect vehicle pass-by from road-side traffic noise measurements. These models may then be implemented as additional features to sensors in a noise sensor network, allowing for the sensors to assist in both the calculation as well as the validation of noise simulations such as those performed in [10].

6. ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge the financial support from the Centre for ECO^2 Vehicle Design (Vinnova Grant 2016-05195), Formas (Grant 2021-01532), Stockholm Region (Grant 2021-0317), the Richert Foundation (Grants 2018-00472 and 2021-00697), and Digital Futures at KTH.

7. REFERENCES

[1] S. Kephalopoulos, M. Paviotti, and F. Anfosso-Lédée, "Common noise assessment methods in europe (cnossos-eu)," 2012.

[2] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ." `https://www.openstreetmap.org`, 2023.

[3] "minidsp umik-1 product datasheet." `https://www.minidsp.com/images/documents/ProductBrief-Umik.pdf`. Accessed: 2023-04-28.

[4] T. Heittola, E. Çakır, and T. Virtanen, "The machine learning approach for analysis of sound scenes and events," *Computational Analysis of Sound Scenes and Events*, pp. 13–40, 2018.

[5] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[6] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 171–175, IEEE, 2015.

[7] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1128–1132, IEEE, 2016.

[8] T.-W. Su, J.-Y. Liu, and Y.-H. Yang, "Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 791–795, IEEE, 2017.

[9] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning for weakly-labeled semi-supervised sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, IEEE, 2020.

[10] S. Baclet, S. Venkataraman, R. Rumpler, R. Billsjö, J. Horvath, and P. E. Österlund, "From strategic noise maps to receiver-centric noise exposure sensitivity mapping," *Transportation Research Part D: Transport and Environment*, vol. 102, p. 103114, 2022.