TRAFFIC COUNTING SYSTEM LEVERAGED WITH A NON-SUPERVISED COUNTING APPROACH

Technical Report

Erwann Betton-Ployon, Abbes Kacem[1]*, Jérôme Mars*[2]

[1] ACOUSTB. 38400 Saint-Martin-d'Hères - France
{erwann.betton-ployon, abbes.kacem}@egis-group.com
[2] Universite Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-Lab. 38000 Grenoble - France
jerome.mars@gipsa-lab.grenoble-inp.fr

ABSTRACT

To face the challenges of urban mobility optimisation, safety and disturbance reduction, traffic monitoring flourishes around anthropized areas. Acoustic monitoring can provide a cost-effective traffic counting system, besides using it as a noise monitoring process. One would expect a traffic monitoring system to identify direction and vehicle type while counting pass-bys on audio segments. Main difficulties are related to the variety of sound landscapes and sources near roadways. Generalisation among recording sites is delicate, and the accuracy depends on the amount of labelled data available per site. In this work, we introduce a non-supervised traffic counting algorithm to complement the existing supervised models. Our traffic counting algorithm uses the recording site metadata to estimate a standard GCC-Phat mask for any pass-by. This mask is applied on the cross-correlation signal of the 4 audio channels, permitting a pass-by detection with direction identification. This information is transmitted to the supervised model, which eventually refines its initial output. The addition of our algorithm counting estimation is highly effective on sites with few available labelled data. A significant RMSE reduction is observed when total duration of real labelled data is inferior to 2 hours.

Index Terms— acoustic vehicle counting, acoustic signal cross-correlation, audio signal analysis

1. INTRODUCTION

For the last decades, most nations have gone through a quick industrialisation and urbanisation process. New populations have undergone profound changes in their living environments. As a side effect, an increasing noise exposure has been observed for these populations. Hence, various institutes communicated on the public health issue that high noise exposure represents [1]. Several studies showed that overexposure is a risk factor for cardio-vascular diseases or diabetes, besides obviously different levels of hearing loss [2]. Consequently, regulation has evolved to better monitor and control sound exposure in different contexts.

A difference has been made between the major sound landscapes or sources. In that respect, traffic noise is the most harmful noise source, according to an European Environment Agency report [1]. Thus, traffic noise monitoring is a key component of the battle against noise overexposure. Furthermore, data gathered by acoustic monitoring systems can also benefit other domains, such as helping urban facilities development or supervising areas of interest [3]. Finally, the use of acoustic monitoring systems offers a significant costs reduction, as collected data are lighter and easier to process than videos, and less intrusive than magnetic or vibration sensors.

Further research has been conducted and several innovative traffic monitoring systems have been proposed [4, 5, 6]. They explore either signal processing or deep learning to solve this task.

As part of the 2024 DCASE (Detection and Classification of Acoustic Scenes and Events) Challenge [7], a baseline model is shared by the organizers to serve as a basis to improve [8, 9]. The objective is to count vehicles on 1-minute long audio segments. More precisely, pass-bys are labelled according to the type of vehicle (car or commercial vehicle) and direction (left or right).

The proposed model is a CRNN that uses both Generalized Cross Correlation with Phase Transforms (GCC-Phat) and learnable Gabor filterbank representations to count pass-bys per type and direction. [8] shows that the model requires fine-tuning for each studied site. Our proposition draws on this model, and aims at improving its behaviour regarding sites with few labelled data. The goal is to provide a higher accuracy in such sites while keeping its usual behaviour when a large dataset is used for training. For this purpose, we choose to include a non-supervised vehicle counting algorithm, that relies on Generalized Cross-Correlation with Phase Transform (GCC-Phat).

When given two simultaneous records from the same site, GCC-Phat is able to mark moving sound sources and their direction of arrival. First, we estimate masks for GCC-Phat resulting from vehicle pass-bys given their direction, speed and distance between source and microphone. Then, a digital image correlation is performed between the estimated mask and the GCC-Phat computed from real audio records. Each pass-by is identified by a correlation local maximum, assuming it respects our assumptions about vehicle speed, direction and distance from microphone.

According to the recording site environment, criteria (threshold, width) on correlation peak selection are adjusted. These successive steps form a proper vehicle counting algorithm per direction. This information eventually permits a final correction of the model prediction, by checking the coherence between the algorithm and supervised models outputs for vehicle counting in each direction. If both predictions diverge, a weighting average is performed, with a β coefficient that depends on the amount of data used for training the supervised model.

In this technical report, section 2 will detail the methodology, regarding the supervised model as well as the vehicle counting al-

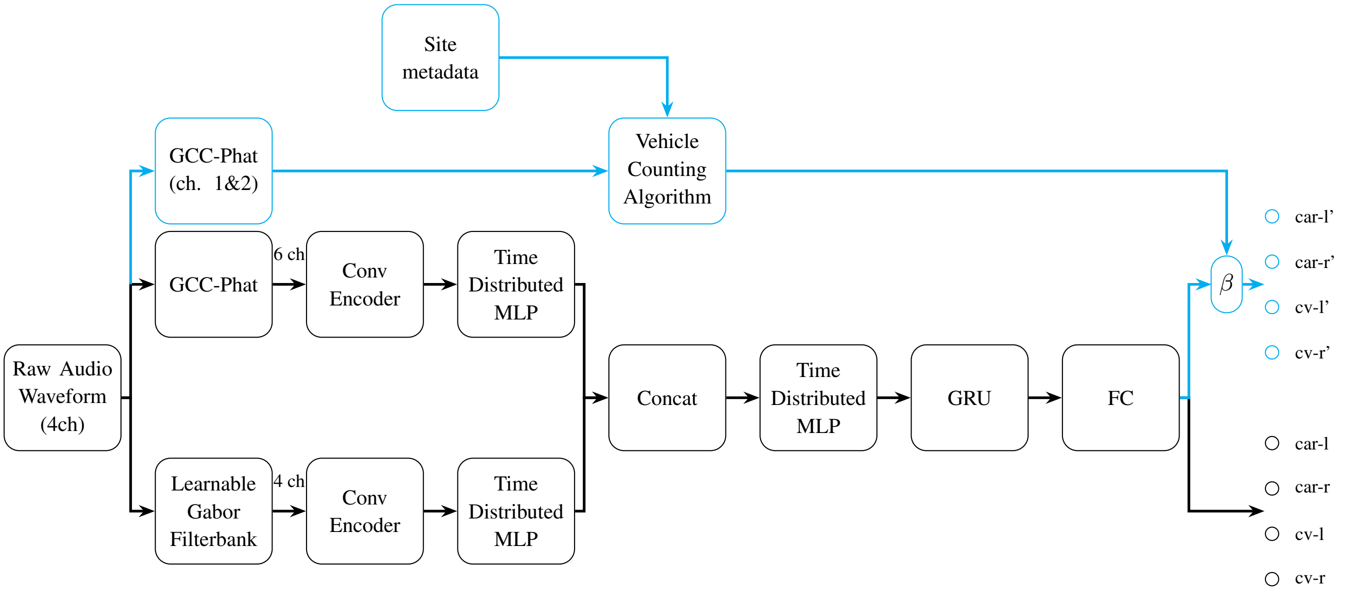

Figure 1: Architecture of the proposed model. In black, the initial baseline model, computing in parallel GCC-Phat between pairs of channels and learnable Filterbank with Gabor filters. Encoded, time-distributed and concatenated features go through a Gated Recurrent Unit (GRU) and a fully-connected (FC) layer to regress the number of vehicles per type (car, CV) and per direction (left-to-right (l), right-to-left (r)) [8]. In cyan, the added units, with a vehicle counting algorithm which estimates the number of vehicles per direction thanks to GCC-Phat on channels 1 & 2 and site metadata (speed limit, distance to street side). The β unit represents a weighting coefficient between the algorithm estimation and the model FC output.

gorithm. In section 3 we will focus on the whole system evaluation, to compare with the initial model behaviour. Finally, section 4 provides a conclusion of this work, with further perspectives for improvement.

2. METHODOLOGY

Most recent traffic counting systems are supervised, meaning that neural networks use labelled data during a first training phase [5, 6]. The network learns to correctly interpret inputs and predict labels thanks to the data at its disposal. The more the model gets representative data, the better its accuracy will be when applied on new unseen data.

Regarding this specific task, traffic noise largely differs from one site to another due to various parameters (vehicle speed, traffic density, reverberation, distance to street side, etc.). This impels us to separate data per site, and train the model on different data for each site. However, data distribution between site is far from uniform : multiple days of signal are available for one site whereas three other have a total recording duration inferior to 2 hours.

2.1. System principle

This is the reason why we choose to complement the existing supervised model with a non-supervised model, able to correct the result for sites presenting a lack of data. In this subsection, both units will first be described before presenting the connection unit which gathers both outputs.

2.1.1. Supervised model architecture

The supervised model was designed and proposed by [8] in the 2024 DCASE Challenge context. Its architecture is depicted by the black cells in fig. 1.

It consists of a convolutional recurrent neural network (CRNN), with two different branchs for treating direction and vehicle type. Starting from the 4-channels raw audio waveform, the first branch performs Generalized Cross-Correlation with Phase Transform (GCC-Phat) on each of the 6 possible pairs of channels. The outcome pass through a convolutional encoder: two Conv2D layers with 32 filters and one layer with 64 filters, each with a kernel size of (3,3) and a stride of 2 in both dimensions. Then, the Time-Distributed MLP unit consists of two time-distributed fully-connected layers with 128 neurons each. In the second branch, the spectrogram of each channel is filtered via a learnable Gabor filterbank with 96 channels. The output passes through two supplementary units (Convolutional encoder and Time-distributed MLP), with the same initial parameters and filters as in the first branch.

Both branches resulting features are concatenated and passed to another time-distributed block with three layers of 128 neurons. Temporal dependencies are treated thanks to the following Gated Recurrent Unit (GRU). GRU is composed of two layers with 128 neurons each. Finally, the fully-connected (FC) layer acts as a regression layer, to provide a 4-element long scalar vector, providing a prediction for amount of pass-bys per direction and per type. This supervised model presents a total of 506 K trainable parameters.

2.1.2. Vehicle Counting Algorithm

The proposed Vehicle Counting Algorithm (VCA) comes in as an independent branch, whose results are used at the very end of the

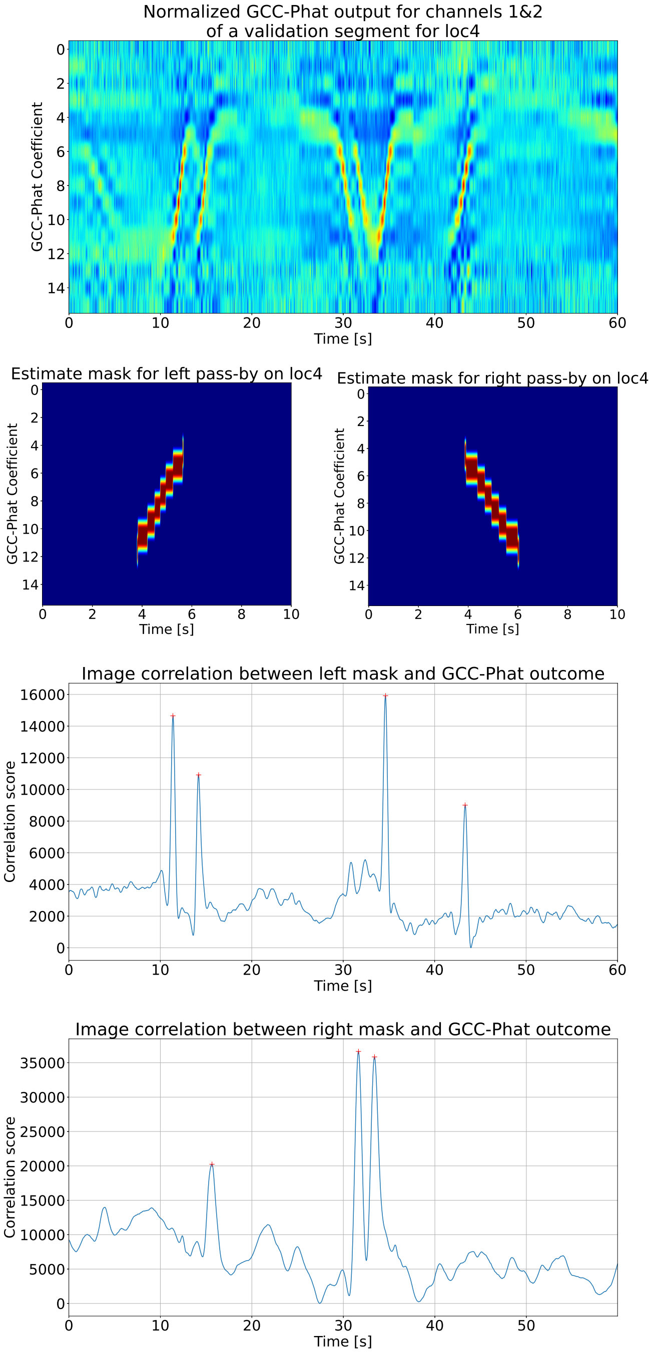

Figure 2: Process of Vehicle Counting Algorithm for estimating the amount of left pass-bys on a 60s-long audio segment. In the two last graphs, red markers stand for detected pass-bys in each direction.

prediction process. Its integration in the proposed system is illustrated in figure 1. A first unit computes GCC-Phat on the first and second channels of the audio input (1^{st} graph of Fig. 2). 16 GCC coefficients are used, with a 64-ms window size and a 32-ms hop length. VCA uses the GCC-Phat outcome and recording site metadata to count vehicles per direction.

The first step of the VCA consists in generating an estimate mask for a typical pass-by according to the site metadata (2^{nd} and

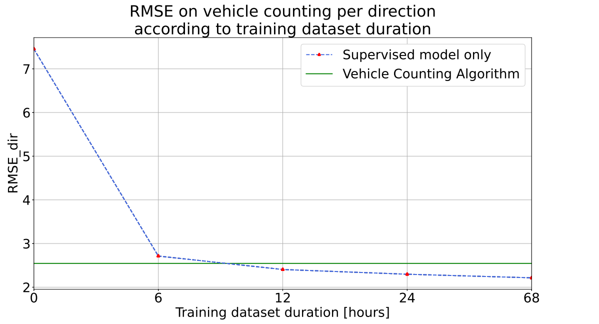

Figure 3: Evolution of RMSE for baseline model and VCA according to training dataset size on site 3.

3^{rd} graphs of Fig. 2). Speed limit and distance to street side indeed impact the shape of a pass-by on the GCC-Phat outcome. Thus, both information are retrieved to generate an adequate mask for each different recording site. The generated mask is a 2D matrix of size $(16 \times L)$, corresponding to the 16 GCC coefficients and L the number of temporal frames.

For a 60s-long audio segment, the GCC-Phat outcome with the described parameters has a 16×1874 size. The correlation between the GCC-Phat outcome and the generated mask provides a feature enabling an accurate pass-by detection (4^{th} and 5^{th} of fig. 2). Each pass-by in a given direction corresponds to a local maximum in the correlation evolution. Additional criteria may be applied to sharpen the pass-by detection. These criteria (threshold, prominence, distance and width) are defined according to both site metadata and empirical findings. On last two graphs of figure 2, detected pass-bys from left and right are identified by red markers.

This peak detection process allows a pass-by counting per direction on the given audio segment. At the same time, the supervised model provides a prediction per site and per vehicle type for the same segment. There now remains to take the final decision using both information.

2.1.3. *Connection between both units*

As written in introduction to this section, a supervised model accuracy highly depends on the amount of representative data available for training. Figure 3 shows RMSE on vehicle counting per direction for various training dataset duration on site 3. Regarding this site, we wee that RMSE is lowered if training dataset is inferior to 6 hours whereas a duration longer than 12 hours favours the supervised model. Thus, we choose to introduce an adaptive weighting coefficient (Eq. 1) on the given inputs while making the final decision.

$$\beta_i = 1 - \min\left(\frac{N_i}{1440}; 1\right) \tag{1}$$

with N_i the amount of training segments for site i.

With β_i defined this way, we have $\beta_i \in [0, 1]$. With few available data, β_i is close to 1. With enough data (N_i close to or superior to 1440), β_i is close to 0. Thus, the β_i coefficient is applied to the VCA counting estimation, whereas a $(1 - \beta_i)$ coefficient is applied to the model output, as described by equation 2.

$$\begin{cases} \text{car\_l}' &= \frac{\text{car\_l}}{\text{car\_l+cv\_l}}\left(\beta_i \times \text{estim\_l} + (1-\beta_i)(\text{car\_l}+\text{cv\_l})\right) \\ \text{car\_r}' &= \frac{\text{car\_r}}{\text{car\_r+cv\_r}}\left(\beta_i \times \text{estim\_r} + (1-\beta_i)(\text{car\_r}+\text{cv\_r})\right) \\ \text{cv\_l}' &= \frac{\text{cv\_l}}{\text{car\_l+cv\_l}}\left(\beta_i \times \text{estim\_l} + (1-\beta_i)(\text{car\_l}+\text{cv\_l})\right) \\ \text{cv\_r}' &= \frac{\text{cv\_r}}{\text{car\_r+cv\_r}}\left(\beta_i \times \text{estim\_r} + (1-\beta_i)(\text{car\_r}+\text{cv\_r})\right) \end{cases} \tag{2}$$

2.2. Training process

Aside from its architecture, a supervised model also relies on an adequate training process. The amount of labelled data available highly impacts the choice of hyperparameters and overall training strategy. In [8], the authors introduce synthetic data to address lack of data for several sites. Model is first pre-trained on synthetic data before fine-tuning on real data. The process of generating synthetic data is described in section 2.2.1.

2.2.1. Data synthesis

The open source `pyroadacoustics` [9] simulator is used to generate data. This simulator is designed to synthesize noise from individual pass-bys on a road. The vehicle modelling relies on a mixture of road/tire interaction noise (generated by Harmonoise model [10]) and engine noise (produced by Baldan model [11]). These signal mixtures are used as inputs to the `pyroadacoustics` simulator, generating several different pass-by events. Finally, 60s-long segments are created using these individual pass-bys, site traffic density and proportion between cars and commercial vehicles.

2.2.2. Hyperparameter definition

To stay in line with the initial model approach [8], pre-training is performed on 24 hours of synthetic data, before fine-tuning with the available real-data per site. Model training is performed using a mean squared error loss, Adam optimizer and a batch size of 64. Learning rate for pre-training phase is equal to 10^{-3}. During fine-tuning, learning rate will be reduced to 10^{-4}.

3. EVALUATION

3.1. Data description

As established earlier, our model is evaluated on 6 different recording sites, in Germany and United States. A total of 264 hours of recording were used for the training, validation and test sets. Table 1 describes the amount of available data, and various properties for each of the 6 recording sites.

These properties can be viewed in parallel with the table 2, presenting error metrics per site for baseline model, VCA only and the proposed model. First, we look at the RMSE_dir, which is the Root-Mean-Square Error on the estimated pass-by counting for each direction. Vehicle type is ignored in the first instance, as VCA only differentiates direction. VCA is better than the baseline model on the 3 sites with fewer available data. However, baseline lowers RMSE_dir for the 3 other sites. This confirms the trend observed in figure 3, where baseline outperforms VCA when training dataset duration is longer than 9 hours.

Table 1: Site metadata and amount of recorded data.

Site	**Training set duration (h)**	**Max traffic density (vehicle/hour/lane)**	**Speed limit (km/h)**
loc1	20.9	1000	100
loc2	0.9	500	50
loc3	68.8	500	50
loc4	0.3	400	50
loc5	1.6	140	40
loc6	29.0	900	90

Table 2: Error metrics evaluated on the validation dataset of each site. RMSE is computed regarding direction only on the baseline and VCA predictions, then regarding direction and type on the baseline and proposed model predictions.

Site	**RMSE_dir (baseline)**	**RMSE_dir (VCA)**	**RMSE (baseline)**	**RMSE (model)**
loc1	4.346	4.620	1.661	1.768
loc2	3.606	2.461	1.987	1.294
loc3	2.215	2.545	0.836	0.839
loc4	1.711	0.964	1.296	0.689
loc5	1.088	0.956	0.609	0.487
loc6	2.703	4.237	1.150	1.155

Regarding the final outcome, the proposed model lowers RMSE compared to baseline on average. A clear improvement is realised on the three sites with fewer data. Otherwise, RMSE stays in the same range, because the β_i coefficient (Eq. 1) is equal to 0, meaning that VCA counting is ignored.

4. CONCLUSION AND FUTURE WORK

In conclusion, we presented a vehicle counting algorithm aimed at complementing an existing supervised model. The algorithm is based on Generalized Cross-Correlation with Phase Transform over channels of a microphone array. It estimates the amount of pass-bys per direction on a given audio segment, using information on the recording site such as speed limit or distance between microphone and street side.

Associated with a supervised model designed to count pass-bys per direction and per vehicle type, our algorithm may refine the initial model predictions. In case of a lack of training data, supervised model is not always accurate on certain sites. Its predictions are then corrected by the implemented vehicle counting algorithm outcome. This addition allows an average 0.474 RMSE decrease for the 3 sites with fewer data, while keeping similar error values for the 3 remaining sites.

In a supervised traffic counting system, the added algorithm is highly effective to cover for a lack of data on specific recording sites. We now aim to improve the algorithm accuracy in dense traffic conditions, in order to provide an efficient system in any conditions. Thus, we could validate a lower standard for the amount of data required to train an accurate traffic counting system.

5. REFERENCES

[1] European Environment Agency, “Noise in Europe 2014”, https://www.eea.europa.eu/publications/noise-in-europe-2014

[2] S. Stansfeld, M. Matheson, “Noise pollution: non-auditory effects on health,” in *British medical bulletin*, vol.68, no. 1, pp. 243–257, 2003.

[3] M. Won, “Intelligent traffic monitoring systems for vehicle classification: A survey,” in *IEEE Access*, vol. 8, pp. 73340–73358, 2020.

[4] A. Severdaks, M. Liepins, “Vehicle Counting and Motion Direction Detection Using Microphone Array,” in *Electronics and Electrical Engineering*, vol. 19, no. 8, Oct. 2013.

[5] S. Djukanovic, Y. Patel, J. Matas et al., “Neural network-based acoustic vehicle counting,” in *Proc. 2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 561–565, Aug. 2021.

[6] N. Bulatovic, S. Djukanovic, “Melspectrogram features for acoustic vehicle detection and speed estimation,” in *Proc. 26th International Conference on Information Technology (IT)*, Feb. 2022.

[7] http://dcase.community/challenge2024

[8] S. Damiano, L. Bondi, S. Ghaffarzadegan et al., “Can Synthetic Data Boost the Training of Deep Acoustic Vehicle Counting Networks?,” in *Proceedings of the 2024 International Conference on Acoustics, Speech and Signal Processing*, Apr. 2024.

[9] S. Damiano and T. van Waterschoot, “Pyroadacoustics: a Road Acoustics Simulator Based on Variable Length Delay Lines,” in *Proceedings of the 25th International Conference on Digital Audio Effects*, pp. 216–223, Sep. 2022.

[10] Hans G. Jonasson, “Acoustical source modelling of road vehicles,” *Acta Acustica united with Acustica*, vol. 93, no. 2, pp. 173–184, 2007.

[11] S. Baldan, H. Lachambre, S. Delle Monache et al., “Physically informed car engine sound synthesis for virtual and augmented environments,” in *Proc. 2015 IEEE 2nd VR Workshop Sonic Interactions Virtual Environments*, pp.1–6, Mar. 2015.