Robust Audio-Based Vehicle Counting in Low-to-Moderate Traffic Flow

Slobodan Djukanović[1], Jiří Matas[1] and Tuomas Virtanen[2]

***Abstract*— The paper presents a method for audio-based vehicle counting (VC) in low-to-moderate traffic using one-channel sound. We formulate VC as a regression problem, i.e., we predict the distance between a vehicle and the microphone. Minima of the proposed distance function correspond to vehicles passing by the microphone. VC is carried out via local minima detection in the predicted distance. We propose to set the minima detection threshold at a point where the probabilities of false positives and false negatives coincide so they statistically cancel each other in total vehicle number. The method is trained and tested on a traffic-monitoring dataset comprising 422 short, 20-second one-channel sound files with a total of 1421 vehicles passing by the microphone. Relative VC error in a traffic location not used in the training is below 2% within a wide range of detection threshold values. Experimental results show that the regression accuracy in noisy environments is improved by introducing a novel high-frequency power feature.**

I. Introduction

Traffic monitoring (TM) is used to collect data about the use and performance of roadway systems. The TM data include estimates of vehicle count, flow rate, vehicle speed, vehicle length and weight, vehicle class and identity via the registration plate [1]. Traffic analysis carried out with the collected TM data enables better use of the roadway systems (e.g., enables drivers to be better informed about traffic congestion and parking possibilities), law enforcement (speeding vehicles, dangerous driving, detection of stolen vehicles), prediction of future transportation needs, and the overall improvement of transportation safety.

Based on the sensor type they use, current TM technologies are classified as intrusive, non-intrusive, or off-roadway [1], [2]. Intrusive sensors are embedded in the road and they include magnetic detectors, vibration sensors, piezoelectric sensors and induction loops. Non-intrusive technologies imply mounting sensors overhead on roadways or roadsides and they include cameras, infrared, ultrasonic, magnetic and acoustic sensors, Laser Infrared Detection and Ranging (LIDAR), and Wi-Fi transceivers. The off-roadway systems use mobile sensors installed on aircrafts or satellites.

Vision-based TM systems have several advantages with respect to the other ones. First, they provide rich vehicle-based information such as visual features, vehicle geometry and precise vehicle path. Second, a single sensor is sufficient for detecting and classifying vehicles in multiple lanes. Third, installing cameras to monitor roadways is cheaper and less disruptive than installing sensors in intrusive systems. However, there are many challenges to implement vision-based TM systems, such as partial occlusion, shadows and illumination variation. In addition, video processing for the TM purpose is computationally expensive and time-consuming [1], [3].

Audio-based TM offers numerous important advantages with respect to the vision-based one [1]. Microphones are less expensive, consume less energy and require less storage space than cameras. They are not affected by visual occlusions and lighting conditions. They are easier to install and maintain, and have low wear and tear. Finally, microphones are less distractive for drivers[1].

This paper addresses vehicle counting (VC) in low-to-moderate traffic flow using one-channel sound. To this aim, we propose a vehicle-to-microphone distance function whose minima correspond to vehicles passing by the microphone. The distance function is predicted using regression and VC is carried out via local minima detection in the predicted distance. We propose to set detection threshold so that the false positives and the false negatives statistically cancel each other in total vehicle count. The obtained counting error in a traffic location not included in training is below 2% within a wide range of detection threshold values. In addition to standard acoustic features, we use a novel feature obtained by integrating a high-frequency spectrum of the considered sound signal. Experimental results show that the regression accuracy in noisy environments is improved by the proposed feature. The method is trained and tested on a TM dataset collected for the purpose of this research.

II. Related work

One of the first attempts to use the sound for automatic estimation of traffic density is presented in [4]. The approach recognizes temporal variations that appear in the power signal when vehicles pass by an observation point. Vehicles are detected based on the state transitions of a hidden Markov model which models global temporal variations of the power signal. Overlapping of the vehicles' sounds is resolved by separately processing channels from a stereo microphone.

In [5] and [6], vehicle detection based on the short-time energy is performed. From the smoothed logarithmic energy,

Slobodan Djukanović was supported by the OP RDE programme under the project International Mobility of Researchers MSCA-IF III at CTU in Prague No. CZ.02.2.69/0.0/0.0/19_074/0016255.

[1]Slobodan Djukanović and Jiří Matas are with Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic {djukaslo,matas}@fel.cvut.cz.

[2]Tuomas Virtanen is with the Audio Research Group, Tampere University, Finland tuomas.virtanen@tuni.fi.

[1]Drivers change their behaviour when they see a camera since they mistake it for a radar.

VC is carried out by detecting peaks via a peak picking algorithm. Spurious peaks due to horns are eliminated by spectral filtering of the sound signal. Features extracted from this signal are used to classify the vehicles into heavy, medium and light categories.

In [7], an unsupervised approach for detecting approaching cars, referred to as Auto++, was proposed. The authors recognized the challenges of the lack of temporal structure of vehicular noise and complex acoustic noise characteristics, which can be further emphasized by low-quality sound recording devices. To address these challenges, a novel robust feature is proposed, top-right frequency, which represents maximum frequency whose power reaches a certain threshold. The added value of this approach is possibility to be implemented on devices with low CPU and memory requirements, such as smartphones, allowing its users to detect oncoming vehicles.

The use of microphone arrays for VC and motion direction estimation is considered in [8] and [9]. The authors in [8] use sound delay in three microphones for these tasks. Small-aperture microphone array is used in [9] and spatial coherence is used to select the useful bands for determining the vehicle direction, counting vehicles and estimating their moving direction.

III. Proposed vehicle counting

A. Assumptions

The road situation considered in our research is limited to one-direction two-lane roads. No assumptions are made regarding the speed of the vehicles, nor it is estimated. The recorded sound is defined as a mixture of the sound produced by vehicles to be counted and the environmental noise, which can include the sound of other vehicles not to be counted (due to vicinity of other roads), weather conditions (no wind to moderate wind, no rain) as well as sound reflections of vehicles to be counted (e.g., due to walls by the road).

B. Dataset collection, preprocessing and splitting

A dataset (audio-video recordings of the road traffic) has been recorded using a GoPro Hero5 Session camera. It was installed on a sidewalk, at a safe distance of at least 0.5 m from the road and at the height of around 1.2 m. Figure 1 presents sample images from the camera in six different two-lane one-direction roads in Prague, Czech Republic. The roads are located away from the city centre, but are within the innermost tariff zones P, 0 and B [10]. We have picked various camera positions (both sides of the road and different angles of the camera with respect to the road) in order not to be sensitive to the actual camera position. The dataset acquisition took place from September to November 2019.

The recorded video material is divided into 422 short, 20-second non-overlapping video clips using the Format Factory application (version 4.9.0). The number of video clips per location is 15, 32, 26, 112, 65 and 172 (Fig. 1, left to right, top to bottom). The sound files (44100 Hz sampling rate, WAV format, 32-bit float PCM) have been extracted from

Fig. 1. Data acquisition at six different locations in Prague.

the video clips using Audacity (version 2.3.2), a free open-source application for recording and editing sound. The total material includes 1421 vehicles passing by the microphone, making an average of 303 vehicles/hour/lane.

Annotation data contain the pass-by-microphone times of all vehicles. We record the relative time from the beginning of the file, measured in seconds with a two-decimal precision. Precise annotations are obtained by visual screening, i.e., by identifying a video frame when a vehicle starts to exit the camera view, which is approximately the moment when it is closest to the microphone. Five vehicle classes are considered in the annotation: motorcycles, cars, vans, buses and trucks.

The dataset is split into two parts. The first part, referred to as VC-PRG-1:5[2], corresponds to the first five recording locations in Fig. 1 and it consists of 250 sound files (20-second segments) with 841 vehicles in total. The second part, referred to as VC-PRG-6, corresponds to the sixth location (last image in Fig. 1), it contains 172 sound files with 580 vehicles in total. VC-PRG-1:5 will be used to evaluate the proposed method via the standard k-fold cross validation where all five locations will be included in both the training and the testing phase, whereas VC-PRG-6 will be used for evaluation on data collected at a location not considered in the training. The dataset is available for download at `http://cmp.felk.cvut.cz/data/audio_vc`.

C. Features

The proposed VC approach is based exclusively on audio cues. When a vehicle is directly in front of the microphone, the energy of the corresponding sound signal will reach a local maximum [4]–[6], i.e., a peak in the energy gives an indication about the presence of a vehicle. However, peaks in energy are also induced by other vehicles (in vicinity of other roads), by wind or some other source.

The features will be calculated from a sound signal $x(n)$, where $n = 1, 2, \ldots, N$ represents discrete time variable and N is the signal length. Description of the features follows, along with the implementation details.

1) Short-term energy: Short-term energy (STE) of $x(n)$ is calculated as

$$\mathrm{STE}(m) = \frac{1}{N_w} \sum_{n=1}^{N_w} x^2(n + (m-1)N_h), \quad m = 1, \ldots, M \tag{1}$$

[2]PRG in VC-PRG-1:5 stands for Prague.

where N_w and N_h represent the window and hop lengths, respectively. The STE length M equals the number of hops, i.e., $M = \lfloor \frac{N-N_w}{N_h} \rfloor + 1$, where $\lfloor \cdot \rfloor$ is the round-down operator. In [5], the authors use Hamming window and an STE smoothing with a lowpass Bessel filter in order to remove high frequency fluctuations.

2) Top-right frequency: Top-right frequency (TRF) [7] represents maximal frequency whose power reaches a predefined threshold *T*, i.e.

$$\text{TRF}(m) = \max\left(f[\![X(m,f) \geq T]\!]\right), \quad f \in [0, f_{max}], \quad (2)$$

where $X(m,f)$ represents a time-frequency power distribution of $x(n)$, f is the frequency, m discrete time variable, and $[\![\cdot]\!]$ is the Iverson bracket defined as

$$[\![S]\!] = \begin{cases} 1, & \text{statement } S \text{ is true} \\ 0, & \text{otherwise.} \end{cases}$$

We will calculate $X(m,f)$ as the spectrogram, the squared modulus of the short-time Fourier transform (STFT) [11].

As a vehicle gets closer to the microphone, the energy across all frequency components rapidly increases, reaching a maximum when the vehicle passes by the microphone. This trend holds also for the signal's TRF (white solid line in Fig. 2). The TRF maxima (dotted lines in Fig. 2) correspond to moments when vehicles pass by the microphone.

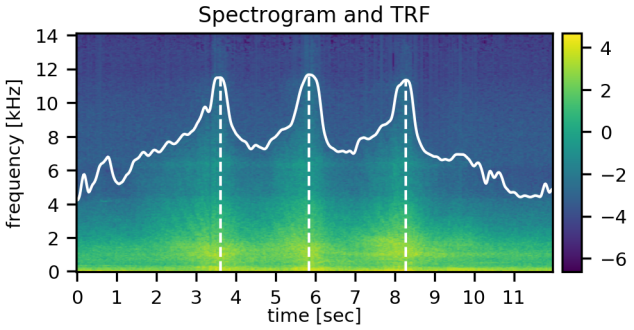

Fig. 2. Spectrogram (log-amplitude scale) and time-varying TRF of the signal.

3) High-frequency power: In this paper, we introduce a new VC feature, calculated as the power of a high-frequency portion of the signal spectrum, and hence referred to as *high-frequency power* (HFP):

$$\text{HFP}(m) = \int_{f=f_{min}}^{f_{max}} X(m,f)df, \quad (3)$$

where f_{min} and f_{max} represent minimal and maximal considered frequencies, respectively. f_{max} can be set to maximum spectral frequency (half of the sampling frequency according to the Nyquist criterion), whereas f_{min} is set so that a low-frequency part of the spectrum, which includes the background noise, is skipped. Figure 3 (top) presents high-frequency (above 6 kHz) portion of the spectrogram of the same signal as in Fig. 2, whereas the bottom plot depicts the proposed HFP feature. The HFP peaks due to vehicles passing by the microphone are much more prominent than the TRF ones.

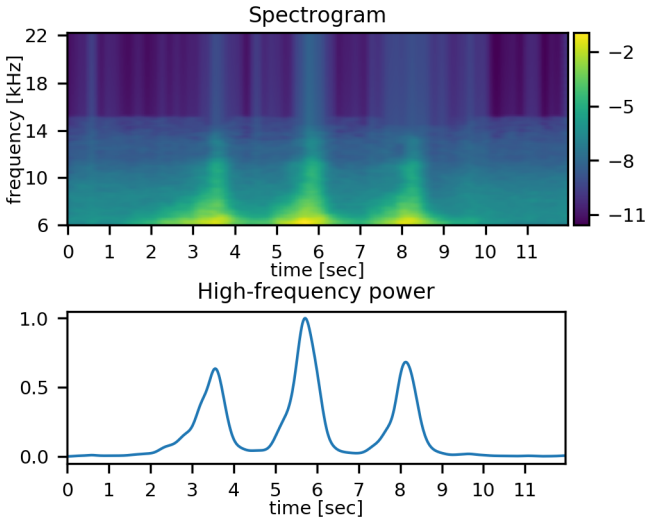

Fig. 3. *Top*: High-frequency (above 6 kHz) portion of the spectrogram (log-amplitude scale). *Bottom*: High-frequency power feature.

4) Log-mel spectrogram: Log-mel spectrogram (LMS) represents the short-term power spectrum of a sound signal projected to a reduced set of frequency bands and converted to logarithmic magnitude. It represents the standard acoustic feature for audio analysis tasks [12].

5) Implementation of features: All the features possess the time dimension, i.e., STE, TRF and HFP are 1-D time functions, whereas LMS is a 2-D function with time as one dimension. The TRF, HFP and LMS features are based on STFT, hence their time dimension depends on the number of window positions in the STFT calculation. In this paper, we use $N_w = 4096$ and $N_h = 0.4N_w = 1638$ samples, which with 20-second audio files sampled at 44100 Hz gives the time-length of all features of 539 samples. In addition, $N_{mel} = 64$ mel bands are used in the LMS calculation, so this feature's shape is 64×539. Hamming window is used in the STFT calculation. In the HFP feature, $(f_{min}, f_{max}) = (6000 \text{ Hz}, 22050 \text{ Hz})$.

Prior to using the STE, TRF and HFP features, we filter them to eliminate high-frequency oscillations. To that end, we apply two successive moving average (MA) filters, first with the length of 11, then of 5 samples. After the filtering, these features are normalized to zero mean and unit variance. Finally, at each time instant n, we do not consider the values of STE, TRF and HFP only at n, but also at K preceding and K following values, i.e., at each time instant, we consider $2K+1$ successive values of these features. The boundary values have been handled by quadratic extrapolation[3]. In this paper, $K = 10$. Therefore, the total number of features is $3(2K+1) + N_{mel} = 127$.

[3]For example, at $n = 1$, the first K out of $2K+1$ values of a feature $F(n)$ are obtained are extrapolating the values $F(1), F(2), \ldots, F(K+1)$.

D. Proposed method

For the VC purpose, we propose to use a function analytically defined as follows:

$$d^{(l)}(t) = \begin{cases} \left|t - T^{(l)}\right|, & \left|t - T^{(l)}\right| < T_d \\ T_d, & \text{elsewhere,} \end{cases} \quad (4)$$

where $T^{(l)}$ represents the moment when the l-th vehicle passes by the microphone and T_d is the distance threshold. Function $d^{(l)}(t)$ has the form of a clipped distance of a vehicle, driven at uniform speed, from the microphone, and is depicted in Fig. 4 (top). Hence, $d^{(l)}(t)$ will be referred to as *clipped vehicle-to-microphone distance* (CVMD) of the l-th vehicle. With multiple vehicles, each vehicle is characterized by a V-shape profile and CVMD takes the general form

$$D(t) = \min\{d^{(1)}(t), d^{(2)}(t), \ldots, d^{(N_v)}(t)\}, \quad (5)$$

where N_v represents the number of vehicles in the corresponding sound signal. CVMD for the signal considered in Figs. 2 and 3 is presented in Fig. 4 (middle), whereas the case when two vehicles are apart by less than $2T_d$ seconds is presented in Fig. 4 (bottom). The value of CVMD is either the distance to the nearest vehicle or T_d.

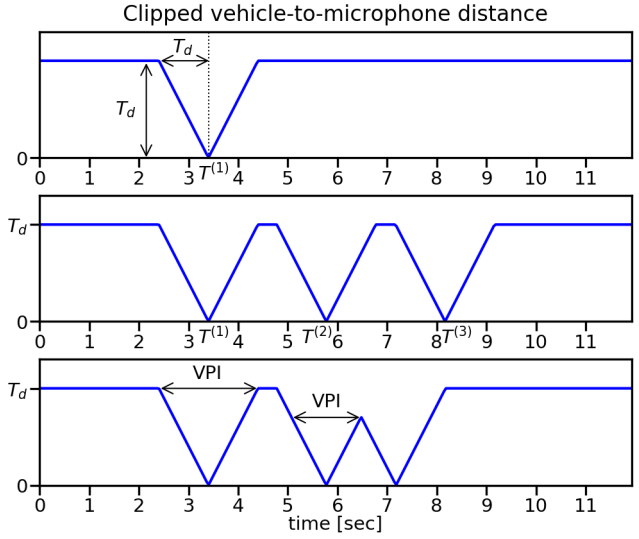

Fig. 4. Clipped vehicle-to-microphone distance (x- and y-axis are in different scale). *Top*: One vehicle. *Middle*: Vehicles are apart by at least $2T_d$ seconds. *Bottom*: Vehicles are less than $2T_d$ seconds apart.

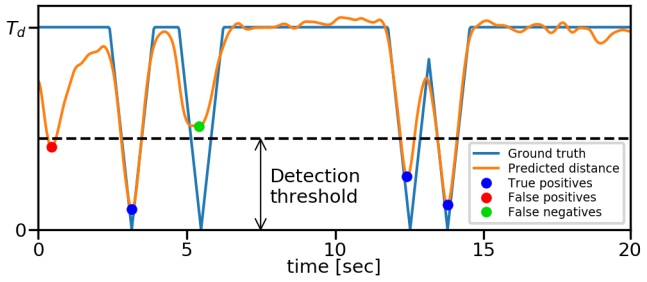

Fig. 5. Classification of local minima in the predicted distance.

The CVMD function allows us to distinguish between the sounds produced by vehicles passing by the microphone and other sounds for which the V-shape distance profile does not make sense. It is predicted using a regression procedure having as input the features described in Section III-C. The regression implementation details follow soon. Figure 5 depicts an example of the regression output (orange line) for a given ground truth CVMD (blue line). Our VC approach is to detect local minima in the predicted distance which surpass a detection threshold (dashed line in Fig. 5). However, not every local minimum surpassing the threshold is associated with a vehicle passing by the microphone (see the red dot in Fig. 5). An additional requirement for a minimum-to-vehicle association is that a local minimum occurs within the corresponding V-shape profile. To that end, we define *vehicle pass-by interval* (VPI) which will represent a basis for deciding whether a local minimum corresponds to a vehicle passing by the microphone (VPI is presented in Fig. 4 (bottom)). Therefore, we define two criteria which will allow us to classify the detected local minima:

C1: local minimum surpasses the detection threshold,
C2: local minimum occurs within a VPI.

A local minimum satisfying both C1 and C2 corresponds to a correctly detected vehicle, or a true positive (TP) (blue dots in Fig. 5). On the other hand, a local minimum satisfying C1 and not C2 corresponds to a false positive (FP) (red dot in Fig. 5), whereas a local minimum satisfying C2 and not C1 corresponds to a false negative (FN) (green dot in Fig. 5). The total number of detected vehicles equals the sum of the TPs and the FPs.

1) Implementation details: In our experiments, for distance threshold we adopt $T_d = 0.75$ s[4]. We apply the support vector machine (SVM) approach for the CVMD regression, more precisely the ε-support vector regression (ε-SVR). The free parameters in the model are C (penalty parameter of the error term) and ε (controls the width of the ε-insensitive zone, used to fit the training data, i.e. determines the accuracy level of the approximated function) [13]. A grid search on C and ε values yielded the optimal values of $C_{opt} = 1$ and $\varepsilon_{opt} = 0.05$. Further reducing the ε value would increase the model complexity and lead to overfit [14].

The SVR is implemented using the `sklearn.svm.SVR` procedure from the scikit-learn library for machine learning in Python. Peak detection is performed using the `scipy.signal.find_peaks` procedure (SciPy v.1.3.0 library for Python), with all parameters set to default values except for the peak prominence set to 0.05. Prior to peak detection, we filter the SVR results to eliminate high-frequency oscillations (three successive MA filters with lengths 7, 5 and 3).

Method implementation in Python is available for download at `http://cmp.felk.cvut.cz/data/audio_vc`.

[4]We did not carry out a detailed analysis of the influence of the T_d value on the method performance. $T_d = 0.75$ is adopted based on a comparison of performances obtained with a few smaller and a few larger values.

IV. Experiment

The proposed method is evaluated by calculating the TP, FP and FN probabilities, denoted as p_{TP}, p_{FP} and p_{FN}, respectively. These probabilities are calculated versus detection threshold varying from 0 to T_d in steps of $\Delta T_d = 1\% T_d$. In addition to these probabilities, we will calculate *normalized area under the curve* (NAUC) p_{TP} as follows:

$$\text{NAUC} = \frac{\sum_{i=0}^{I_{max}-1} p_{\text{TP}}(i\Delta T_d)\Delta T_d}{T_d}, \tag{6}$$

where $p_{\text{TP}}(i\Delta T_d)$ represents the value of p_{TP} obtained for detection threshold $i\Delta T_d$, whereas I_{max} is the number of detection thresholds, in our case 100. Clearly, the bigger NAUC the better, and NAUC $= 1$ is the optimal value.

The proposed method requires selecting the detection threshold. The natural choice would be a point where p_{FP} and p_{FN} coincide since then, in statistical sense, the FPs and the FNs cancel each other so that the total number of detected vehicles equals the true number of vehicles. Therefore, another evaluation metric will be introduced, *equal false probabilities* (EFP), as a value EFP $= p_{\text{FP}} = p_{\text{FN}}$. Having in mind the discrete nature of detection threshold, EFP will be calculated as

$$\text{EFP} = p_{\text{FP}}(I_{min}\Delta T_d), \tag{7}$$

where

$$I_{min} = \underset{i}{\text{argmin}} \; |p_{\text{FP}}(i\Delta T_d) - p_{\text{FN}}(i\Delta T_d)|. \tag{8}$$

As opposed to NAUC, the smaller EFP the better. Ideally, EFP $= 0$.

Finally, let us define *relative VC error* (RVCE) as

$$\text{RVCE} = \frac{|N_v^{true} - N_v^{est}|}{N_v^{true}} \times 100\,[\%], \tag{9}$$

where N_v^{true} and N_v^{est} represent the true and the estimated number of vehicles in the considered dataset.

A. Regression with all features

Figure 6 depicts the TP, FP and FN probabilities obtained after regression with all features. With low detection thresholds, criterion C1 is rarely met, resulting in low p_{TP} and high p_{FN}. An opposite situation is obtained for high detection thresholds, when majority of detected local minima surpass the threshold. In that case, p_{FP} also rises since numerous false minima will also surpass the threshold. The NAUC and EFP values are presented in Table I (Setup I). In addition to EFP, we report the minimal absolute difference between p_{FP} and p_{FN}, i.e.

$$\Delta_{\text{EFP}} = |p_{\text{FP}}(I_{min}\Delta T_d) - p_{\text{FN}}(I_{min}\Delta T_d)|, \tag{10}$$

with I_{min} defined in (8).

The TP probability is high for a wide range of detection threshold values. For example, p_{TP} exceeds 90% and 94% for thresholds above $46\% T_d$ and $71\% T_d$, respectively. Detection threshold corresponding to EFP ($p_{\text{FP}} = 5.52\%$ and $p_{\text{FN}} = 5.45\%$) equals $78\% T_d$ and is shown in Fig. 6 (dashed line). For this threshold value, $p_{\text{TP}} = 94.55\%$.

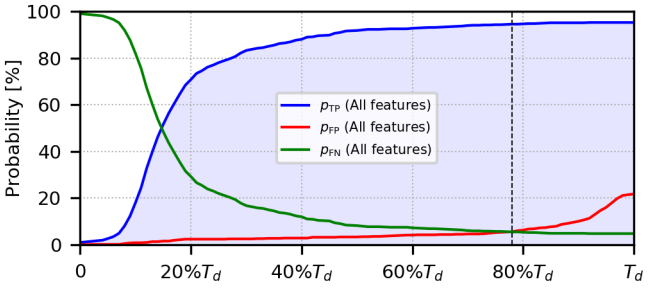

Fig. 6. TP, FP and FN probabilities. Dashed line corresponds to detection threshold where $p_{\text{FP}} = p_{\text{FN}}$.

B. Regression with feature combinations - Ablation study

To see how the exclusion of features affects the regression performance, we repeat the experiment from Section IV-A with different feature combinations. More precisely, we consider four sets of features, each obtained by leaving out exactly one feature from the feature set. The corresponding probability curves are presented in Fig. 7, whereas the NAUC and EFP metrics are given in Table I (Setup II). Clearly, the exclusion of LMS (dash-dot line in Fig. 7 and the last combination in Setup II in Table I) affects performance the most. There are no significant differences in performances of the other three combinations. TRF+HFP+LMS is characterized by the biggest NAUC of all feature combinations in Table I.

Having identified LMS as the most critical feature, we repeat the experiment with i) two-feature combinations, one of them being LMS, and ii) LMS alone. The results are presented in Fig. 8 and Table I (Setups III and IV). The best performance in terms of both NAUC and EFP is obtained with HFP+LMS. Moreover, this feature combination outperforms all other combinations in Table I in terms of EFP. Finally, the NAUC performance of LMS alone is notably worse than those of other combinations, which can be also seen in Fig. 8 (solid line).

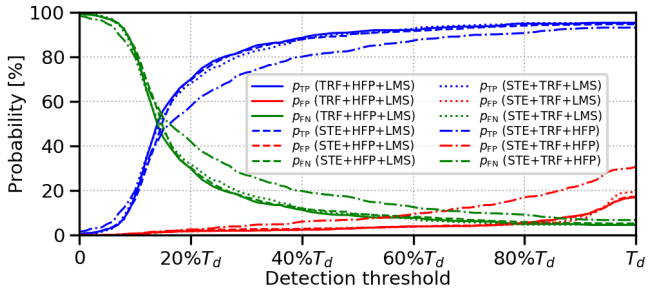

Fig. 7. TP, FP and FN probabilities for three-feature combinations.

C. Method generalization

In this experiment, the method is trained on VC-PRG-1:5 and tested on VC-PRG-6. The location corresponding to VC-PRG-6 differs from those of VC-PRG-1:5 in a road-noise barrier erected on the other side of the road, which causes undesirable acoustic reflection. We consider all features and two best performing combinations (TRF+HFP+LMS and

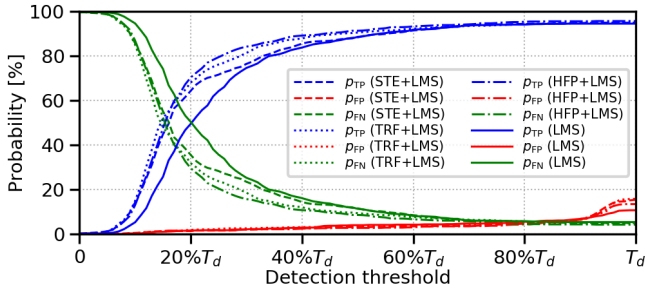

Fig. 8. TP, FP and FN probabilities for combinations of LMS with other features and LMS alone.

TABLE I

NAUC AND EFP METRICS FOR THE VC METHOD TRAINED AND TESTED ON VC-PRG-1:5. $\Delta_{\text{EFP}} = \min|p_{\text{FP}} - p_{\text{FN}}|$.

Setup #	Features	NAUC	EFP [%] (Δ_{EFP} [%])
I	All features	0.773	5.52 (0.07)
II	TRF+HFP+LMS	**0.775**	5.09 (0.06)
	STE+HFP+LMS	0.766	5.68 (0.23)
	STE+TRF+LMS	0.770	4.99 (0.10)
	STE+TRF+HFP	0.721	10.91 (0.06)
III	STE+LMS	0.748	5.49 (0.09)
	TRF+LMS	0.764	5.40 (0.02)
	HFP+LMS	0.772	**4.43** (0.09)
IV	LMS	0.719	5.75 (0.07)

HFP+LMS) from Table I. The results are presented in Fig. 9 and Table II. As in Table I, TRF+HFP+LMS provides the biggest NAUC, whereas HFP+LMS provides the smallest EFP.

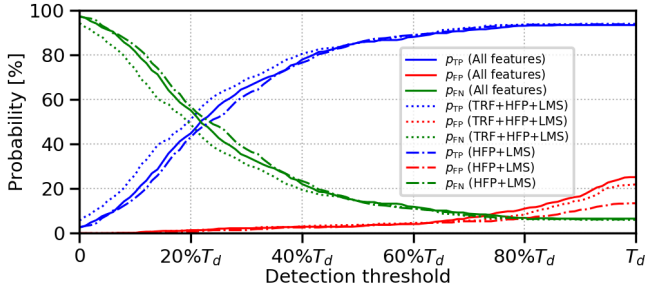

Fig. 9. TP, FP and FN probabilities for the case when different datasets are used for training and testing.

With respect to the first experiment, detection thresholds corresponding to EFP changed from $78\%T_d$ to $74\%T_d$ (all features), $78\%T_d$ to $75\%T_d$ (TRF+HFP+LMS) and $81\%T_d$ to $79\%T_d$ (HFP+LMS). In that sense, HFP+LMS is characterized by the most stable behaviour.

In Fig. 10, we present the RVCE values for detection thresholds within $[50\%T_d, T_d]$. In terms of RVCE, HFP+LMS (red line) significantly outperforms the other two combinations. Namely, its RVCE is below 2% within $[74\%T_d, 85\%T_d]$, as opposed to $[70\%T_d, 76\%T_d]$ (all features, blue line) and $[72\%T_d, 80\%T_d]$ (TRF+HFP+LMS, green line). Arrows in Fig. 10 indicate the threshold ranges where RVCE $< 2\%$. For the EFP thresholds obtained in

TABLE II

NAUC AND EFP METRICS FOR THE VC METHOD TRAINED ON VC-PRG-1:5 AND TESTED ON VC-PRG-6.

Features	NAUC	EFP [%] (Δ_{EFP} [%])
All features	0.702	8.45 (0.34)
TRF+HFP+LMS	**0.729**	6.72 (0.17)
HFP+LMS	0.695	**6.55** (0.17)

the first experiment, RVCE $= 2.76\%$ (all features), 1.21% (TRF+HFP+LMS) and 0.52% (HFP+LMS). Difference in performance is much more striking with very high thresholds. For example, for the highest value of threshold, RVCE $= 18.62\%$ (all features), 15.86% (TRF+HFP+LMS) and 7.24% (HFP+LMS).

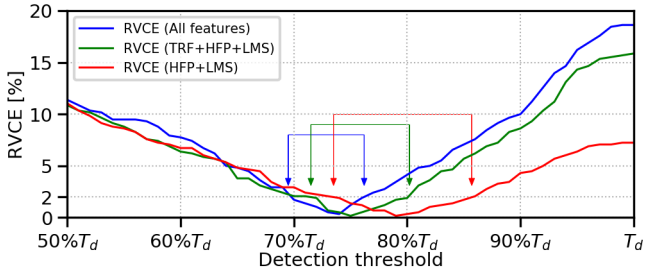

Fig. 10. Relative VC error for the case when different datasets are used for training and testing. Arrows delimit intervals where RVCE is below 2%.

This experiment stresses the importance of introducing the HFP feature in the case of higher environmental noise. By analyzing the behaviour of the 1-D features (STE, TRF and HFP), we may conclude that HFP is much more robust to environmental noise than STE and TRF, i.e., it produces less false peaks than the latter two. This is illustrated in Fig. 11, which depicts these features of two sound files recorded at the same location, one with low (top plot) and the other with high environmental noise (bottom plot). In addition to being robust to environmental noise, the HFP peaks are narrower than those of STE and TRF, which provides a capability of improved separation of close vehicles.

V. CONCLUSIONS

We proposed a method for sound-based vehicle counting in low-to-moderate traffic flow. The method is based on a novel vehicle-to-microphone distance function which allows us to distinguish between the sounds due to vehicles passing by the microphone and all other sounds. The proposed distance is predicted using standard acoustic features and one novel feature which improves prediction robustness in noisy environments. The robustness is manifested in a very low counting error within a wide range of detection threshold values when testing is performed for a traffic location not considered in training.

The future research will aim to identify and resolve error-causing situations in the proposed method, such as sound occlusion due to vehicles passing by the microphone simultaneously. We will also consider other regression approaches, such as recurrent neural networks, which enable

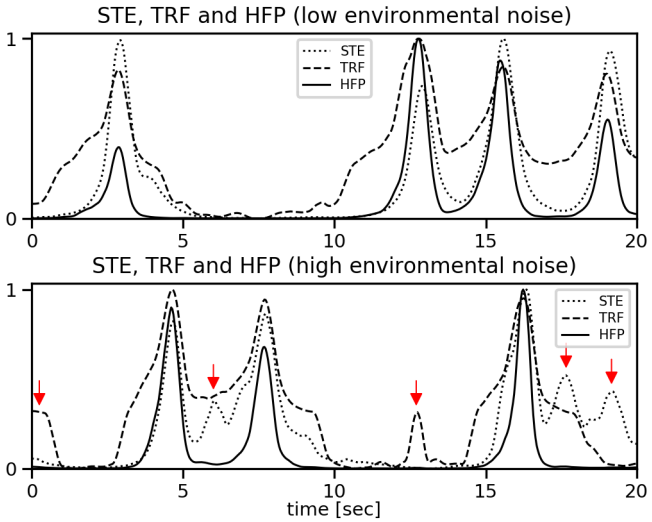

Fig. 11. STE, TRF and HFP features in low-noise (top plot) and high-noise environment (bottom plot). Red arrows in the bottom plot indicate the positions of acoustic-noise (false) peaks.

accurate modelling of the sequential data. Finally, other traffic monitoring issues, such as vehicle speed estimation, will be addressed.

REFERENCES

[1] M. Won, "Intelligent traffic monitoring systems for vehicle classification: A survey," *arXiv preprint arXiv:1910.04656*, 2019.

[2] W. Balid, H. Tafish, and H. H. Refai, "Intelligent vehicle counting and classification sensor for real-time traffic surveillance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 6, pp. 1784–1794, 2017.

[3] B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1114–1127, 2008.

[4] J. Kato, "An attempt to acquire traffic density by using road traffic sound," in *Proceedings of the 2005 International Conference on Active Media Technology, 2005.(AMT 2005).* IEEE, 2005, pp. 353–358.

[5] J. George, L. Mary, and K. Riyas, "Vehicle detection and classification from acoustic signal using ANN and KNN," in *2013 international conference on control communication and computing (ICCC).* IEEE, 2013, pp. 436–439.

[6] J. George, A. Cyril, B. I. Koshy, and L. Mary, "Exploring sound signature for vehicle detection and classification using ANN," *International Journal on Soft Computing*, vol. 4, no. 2, p. 29, 2013.

[7] S. Li, X. Fan, Y. Zhang, W. Trappe, J. Lindqvist, and R. E. Howard, "Auto++: Detecting cars using embedded microphones in real-time," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 70, 2017.

[8] A. Severdaks and M. Liepins, "Vehicle counting and motion direction detection using microphone array," *Elektronika ir Elektrotechnika*, vol. 19, no. 8, pp. 89–92, 2013.

[9] X. Zu, S. Zhang, F. Guo, Q. Zhao, X. Zhang, X. You, H. Liu, B. Li, and X. Yuan, "Vehicle counting and moving direction identification based on small-aperture microphone array," *Sensors*, vol. 17, no. 5, p. 1089, 2017.

[10] Prague integrated transport. [Online]. Available: https://pid.cz/en/tariff-details/tariff-zones/

[11] LJ. Stanković, I. Djurović, S. Stanković, M. Simeunović, S. Djukanović, and M. Daković, "Instantaneous frequency in time–frequency analysis: Enhanced concepts and performance of estimation algorithms," *Digital Signal Processing*, vol. 35, pp. 1–13, 2014.

[12] R. Serizel, V. Bisot, S. Essid, and G. Richard, "Acoustic features for environmental sound analysis," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 71–101.

[13] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[14] D. Mattera and S. Haykin, "Support vector machines for dynamic reconstruction of a chaotic system," in *Advances in kernel methods*. MIT Press, 1999, pp. 211–241.