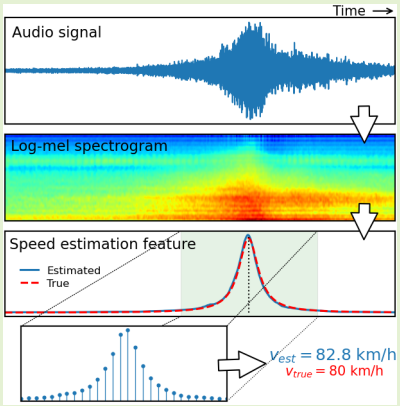
Acoustic vehicle speed estimation from single sensor measurements

Slobodan Djukanović, *Member, IEEE*, Jiří Matas, *Member, IEEE*, and Tuomas Virtanen, *Fellow, IEEE*

***Abstract*— The paper addresses acoustic vehicle speed estimation using single sensor measurements. We introduce a new speed-dependent feature based on the attenuation of the sound amplitude. The feature is predicted from the audio signal and used as input to a regression model for speed estimation. For this research, we have collected, annotated, and published a dataset of audio-video recordings of single vehicles passing by the camera at a known constant speed. The dataset contains 304 urban-environment real-field recordings of ten different vehicles. The proposed method is trained and tested on the collected dataset. Experiments show that it is able to accurately predict the pass-by instant of a vehicle and to estimate its speed with an average error of 7.39 km/h. When the speed is discretized into intervals of 10 km/h, the proposed method achieves the average accuracy of 53.2% for correct interval prediction and 93.4% when misclassification of one interval is allowed. Experiments also show that sound disturbances, such as wind, severely affect acoustic speed estimation.**



***Index Terms*— log-mel spectrogram, neural network, speed estimation dataset, support vector regression, vehicle speed estimation**

I. Introduction

Traffic monitoring (TM) systems collect various traffic data on the use and performance of roadway systems. The data include estimates of vehicle count, speed and class, as well as of vehicle length, weight and identity via registration plates [1]. Based on the collected data, improvements can be made in the performance of roadway systems, transportation safety, law enforcement and prediction of future transportation needs.

Reliable automatic speed detection of moving vehicles is crucial to traffic law enforcement in most countries, and is considered an important tool in decreasing traffic accidents and fatalities. For example, [2] reports that it leads to a reduction of "11% to 44% for fatal and serious injury crashes". Compliance to speed limits is currently monitored with speed enforcement cameras which use Doppler radar, Laser Infrared Detection and Ranging (LIDAR), stereo vision or automatic number-plate recognition. Although these devices usually perform well, they are expensive and thus cannot be widely used.

Acoustic-based TM offers several advantages with respect to other technologies, such as low price, low amount of energy required for operation, low storage space needed, low installation and maintenance costs to name a few [1]. Acoustic-based speed estimation can be divided into approaches based on single microphone measurements [3]–[10] and those based on microphone arrays [11]–[13]. When a single microphone is used, wave propagation effects are exploited to determine the source movements with the following three assumptions i) vehicle is a point source [3], [4], ii) vehicle's sound has stationary characteristics and can be modeled by an autoregressive moving average model [4], and iii) vehicle produces a pure tone [3]. These assumptions, however, are only partially satisfied, which degrades the performance of estimation algorithms based on them when applied to field data [5]. In addition, Cevher *et al.* in [5] argue that signal frequency information in Doppler-based speed estimation [6] is not useful when a single microphone is used. Authors in [7] focus on detecting changes of speed, i.e., they use several machine learning methods (support vector machine, random forests, neural networks) to detect accelerating, decelerating, and maintaining stable speed of a vehicle. Method [8] uses the pass-by sound to classify the speed and gear position of a vehicle. The reported speed classification results, obtained using gradient boosting and correlation matrix optimization, are near perfect (over 99%) when the speed is discretized between 10 and 5 km/h, and very high (over 90%) with smaller discretization intervals. Method [9] estimates the speed using a neural network and a set of features including the engine firing rate (strongest tone of the signal with frequency below 250 Hz), the envelope of the short-time power spectrum of the signal, mel frequency cepstral coefficients and zero-crossing rate. Estimating the speed using sound emissions recorded by an on-board microphone has also been tackled in the literature. For example, in [10], wavelet packet analysis (WPA) is applied

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

Slobodan Djukanović is with University of Montenegro, Faculty of Electrical Engineering, Podgorica, Montenegro (e-mail: slobdj@ucg.ac.me).

Jiří Matas is with Czech Technical University, Faculty of Electrical Engineering, Prague, Czech Republic (e-mail: matas@fel.cvut.cz).

Tuomas Virtanen is with Tampere University, Audio Research Group, Tampere, Finland (e-mail: tuomas.virtanen@tuni.fi).

on sound emissions and the speed is estimated using a neural network fed by norm entropy of subsignals from WPA. The reported average prediction rate is 97.89% with 1.11 km/h mean absolute error and 2.11% relative error.

Microphone-array based approaches exploit the correlation of signals coming from separate microphones. Authors in [11] propose an estimation technique based on the maximum likelihood principle. No assumptions are made regarding the acoustic signal emitted by a vehicle, which has the advantages of bypassing troublesome intermediate delay estimation steps with respect to competing methods. In [12], a nonlinear least squares method for speed estimation is proposed, based on time-delay-of-arrival estimates from multiple microphones. A quasi-Newton method is used to improve the computational efficiency. Estimation of speed and wheelbase of two-axle vehicles is addressed in [13]. The method assumes that the pass-by sound is mainly composed of tyre/road noise. Microphone array considered in [13] contains two microphones. The absolute difference between the true and estimated speeds is below 5 km/h for 75% of all considered vehicle runs.

One of the main challenges to acoustic vehicle speed estimation is a lack of labeled data. Datasets used in experiments in the aforementioned studies are very small. For example, in [5], ten audio recordings (nine different vehicles) were used. In [6] and [11], only seven recordings were used (three cars, a bus and a motorbike). In [9], two different cars were used, four different speeds per car, two recordings per speed. Experiments in [8] used the sound "of an American-built car driving multiple identical laps on a closed parking lot", without specifying car manufacturer, speed and the number of laps. In [13], one recording of 240 seconds, containing 22 and 2 motorbikes, was used. In [10], one vehicle was used, with speeds from 30 km/h to 80 km/h, 1 km/h increment.

This paper addresses acoustic speed estimation using measurements from a single sensor. We propose a method that is able to i) accurately detect passing vehicles and to ii) estimate their speed with an average error of 7.39 km/h. A dataset of 304 audio-video recordings of vehicles passing by the sensor at constant speed is collected, annotated, and made publicly available. It is suitable for both video- and audio-based vehicle speed estimation. To the best of our knowledge, this is the most extensive annotated dataset for vehicle speed estimation.

The dataset is presented in Section II. Speed estimation method is presented in Section III and experimentally verified in Section IV. Section V concludes the paper and gives future research directions.

II. Dataset

We have collected a dataset of on-road recordings of single vehicles passing by the camera. Ten different vehicles were used resulting in a total of 304 audio-video recordings, each one containing a single drive of a single vehicle. The main goals set before compiling the dataset were: (1) recordings should be made in an urban environment, (2) recordings have to be real field ones, (3) vehicles should be as diverse as possible in terms of manufacturer, production year, engine type, power and transmission, (4) all vehicles have to be equipped with the cruise control system, so that speed is maintained stable during the vehicle's pass by.

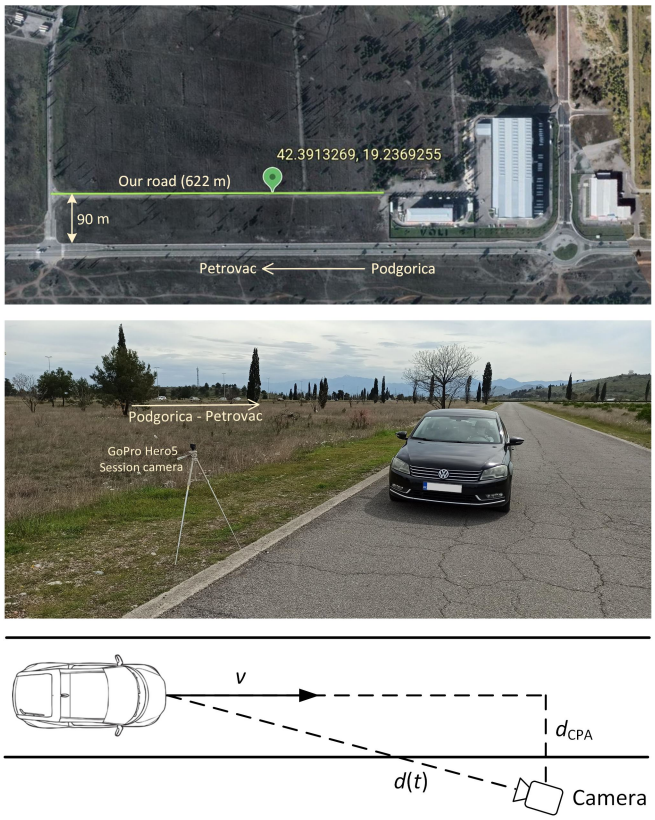


Fig. 1. *Top:* Google map screenshot of the recording site. Our road (green line, **622** m long) is **90** m away from the main road Podgorica-Petrovac (Montenegro). *Middle:* Screenshot of the recording setup. *Bottom:* Vehicle moving at a constant speed v on a straight path. $d(t)$ is the distance between a vehicle and the camera at time instant t, whereas d_{CPA} is the distance at the closest point of approach (CPA).

In the context of acoustic vehicle speed estimation, goals (1) and (2) imply that in addition to the pass-by sound of vehicles used in the experiment (prominent sound source), audio files can contain sounds of other nearby vehicles and environmental sounds (e.g., wind, bird chirps, crickets), which are considered as noise in the estimation.

The dataset (recordings with annotations), referred to as VS10, is available for download at `http://slobodan.ucg.ac.me/science/vse/`. For convenience, we have extracted audio files and provided them separately for download. These audio files represent the material on which our experiments were conducted and results presented in Section IV. More details on preparing the dataset follow.

A. Dataset collection and preprocessing

The dataset was recorded on a local road (green line in Fig. 1 (top)), 622 m long, located 90 m away from the main road Podgorica-Petrovac in Montenegro. This road is selected for the following reasons: i) it is long enough so that stable speeds can be achieved prior to the pass-by instant, ii) it is isolated enough to allow measurements without too many disturbances, and iii) it is close to other roads so that dataset creation goals (1) and (2) are fulfilled.

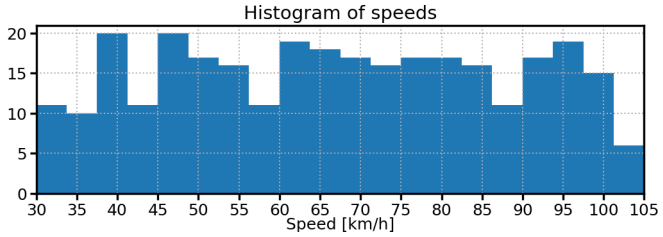


Fig. 2. Histogram of VS**10** speeds calculated at **20** equal-width bins.

For dataset recording, we used a GoPro Hero5 Session camera. It was installed by the road, at a distance of around 0.5 m from the road and at a height of around 1.2 m (see Fig. 1 (middle)). The camera was installed in various positions (both sides of the road and different angles with respect to the road) in order not to be sensitive to the actual camera position[1]. The recording sessions (one session per day) took place from December 2019 to December 2020. Ten vehicles were used, as listed in Table I. The number of recording sessions per vehicle is given in the sixth column of Table I.

The speed of vehicles ranges from 30 to 105 km/h, with the exact values given in the rightmost column in Table I. Below 30 km/h, the cruise control cannot be used with the selected vehicles (for Peugeot 3008, even below 40 km/h). Above 105 km/h, the selected road did not allow for stable and secure measurements. For each vehicle, we have adopted a variable speed step, from 1 to 3 km/h, so that practically all speeds from 30 to 105 km/h are included in the VS10 dataset. Histogram of speeds is given in Fig. 2. The reported speeds are stable[2] at least 3 seconds before and after the pass by. Outside that 6-second interval, minor speed variations are possible.

Each recording in VS10 contains a single drive of a single vehicle. The original recordings were cut into 10-second video files[3] (using the Format Factory application) so that the pass-by instants of vehicles are around the middle of the file length. Each video file is accompanied by an annotation text file which contains the speed of the vehicle and its pass-by-camera instant. We measured the relative time from the beginning of the file, given in seconds with a two-decimal precision. Precise annotations were obtained by visual screening, i.e., by identifying a video frame when the vehicle starts to exit the camera view, which approximately corresponds to the closest point of approach (CPA) (see Fig. 1 (bottom)).

For the purpose of acoustic vehicle speed estimation, we extracted audio files (44100 Hz sampling rate, WAV format, 32-bit float PCM) from the corresponding video files using Audacity, a free open-source application for recording and editing sound. A signal containing the sound of a vehicle passing by the camera is presented in Fig. 4 (top).

The VS10 dataset contains 10 folders, with 304 video files in total. Each folder corresponds to one vehicle, i.e., it contains 10-second video files (MP4 format, full HD resolution, 30 fps) and annotations for that vehicle. The extracted audio files, used in experiments, are also available for download via a separate link. In addition to 304 audio files containing the sound of vehicles passing by the camera, we provide additional 36 audio files (without corresponding video) containing only environmental noise (no vehicles passing by the camera), recorded using the same setup. The additional files are included to improve regression of introduced speed estimation feature.

Naming convention for dataset files includes the vehicle name and the speed. For example, `Peugeot307_79.mp4`, `Peugeot307_79.wav` and `Peugeot307_79.txt` represent the names of video, audio and annotation files, respectively, of Peugeot 307 driven at 79 km/h. Additional no vehicle audio and annotation files are named e.g. `NoCar_021.wav` and `NoCar_021.txt`.

Finally, note that the considered experimental scenario is limited in the sense that drives of single vehicles were recorded. In a large-scale city traffic scenario, several vehicles can simultaneously pass by the microphone and the recorded sound can contain several overlapping components. In that case, sound-based speed estimation cannot be performed without preprocessing the recorded sound, i.e., without proper separation of sound components. This issue is beyond the scope of this paper.

B. Train-validation split

The proposed method will be evaluated using 10-fold cross validation. To that end, we split files in each folder to training and validation files, 80%–20% split. The split procedure is as follows: i) sort the speeds into ascending order, ii) divide the sorted speeds into batches of 5 speeds, iii) randomly select one speed in each batch to be used for validation, the other ones for training. This strategy ensures that low-, medium- and high-speed audio are used in both training and validation. Each folder contains a file `Train_valid_split.txt` with labels *train* or *valid* associated with each audio.

III. Speed Estimation

Our speed estimation approach is based solely on audio obtained from a single microphone (in our case, audio is extracted from video). We introduce a new speed-dependent feature that will be predicted from the input audio (Section III-A). Vehicle speed is estimated via a regression approach having as input the predicted feature (Section III-B).

A. Speed estimation feature

Our feature is based on the amplitude attenuation factor of the sound signal [4]

$$\sigma(t) = \frac{1}{\sqrt{v^2(t_{\text{CPA}} - t)^2 + d_{\text{CPA}}^2}}, \tag{1}$$

where v represents speed, t time variable, t_{CPA} the CPA instant, and d_{CPA} distance at CPA. Attenuation $\sigma(t)$ for speeds $v = [30, 55, 80, 105]$ km/h, $t_{\text{CPA}} = 5$ s and $d_{\text{CPA}} = 1.5$ m is presented in Fig. 3. Observe that the $\sigma(t)$ shapes are close to each other, especially for higher speeds, which renders $\sigma(t)$ unreliable for speed estimation.

[1]Sample video files of each vehicle can be seen at `http://slobodan.ucg.ac.me/science/vse/`.

[2]The speed is maintained stable by the on-board cruise control, all vehicles were equipped with.

[3]Authors in [7] also segment the recorded material into 10-second files.

TABLE I
VS10 VEHICLES AND SPEEDS

Vehicle	Engine type	Power (kW)	Transmission	Prod. year	Record. sessions	Speeds (km/h)
Citroen C4 Picasso	Diesel	88	Manual	2015	1	35, 38, 41, 44, 48, 51, 54, 57, 59, 63, 65, 68, 72, 74, 78, 80, 83, 85, 87, 92, 94, 96, 101
Mazda 3 Skyactive	Petrol	74	Manual	2015	1	30, 33, 35, 38, 40, 43, 45, 47, 50, 52, 55, 57, 60, 62, 64, 67, 70, 72, 75, 79, 81, 84, 86, 88, 90, 92, 94, 96, 99, 101, 103, 105
Mercedes AMG 550	Petrol	350	Automatic	2006	3	30, 33, 35, 38, 40, 42, 45, 47, 50, 52, 55, 58, 60, 62, 65, 67, 70, 73, 75, 78, 80, 82, 85, 87, 90, 93, 95, 98, 100, 105
Nissan Qashqai	Diesel	81	Manual	2018	1	35, 38, 40, 42, 45, 48, 50, 53, 55, 58, 60, 61, 64, 65, 68, 70, 73, 75, 78, 80, 82, 85, 88, 90, 93, 94, 96, 98, 102
Opel Insignia	Diesel	96	Automatic	2010	1	31, 35, 38, 41, 44, 47, 50, 53, 55, 58, 61, 64, 66, 68, 70, 72, 73, 76, 78, 80, 83, 86, 89, 91, 94, 97, 100
Peugeot 3008	Diesel	84	Automatic	2013	2	40, 43, 45, 47, 50, 52, 54, 55, 56, 58, 60, 61, 63, 65, 67, 68, 70, 72, 74, 75, 78, 80, 83, 85, 87, 89, 90, 92, 95, 97, 100
Peugeot 307	Diesel	100	Manual	2007	1	30, 33, 35, 38, 40, 43, 45, 47, 48, 50, 53, 56, 59, 60, 63, 66, 69, 72, 73, 76, 79, 82, 85, 88, 91, 94, 97, 101, 103
Renault Captur	Diesel	66	Automatic	2015	1	30, 33, 36, 38, 40, 41, 44, 46, 47, 48, 50, 52, 56, 58, 60, 63, 66, 68, 70, 72, 76, 78, 80, 83, 86, 88, 90, 92, 94, 97, 98, 100, 102
Renault Scenic	Diesel	96	Manual	2010	2	30, 35, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 57, 60, 62, 64, 66, 68, 70, 71, 72, 74, 75, 77, 80, 82, 84, 86, 87, 90, 91, 94, 95, 98, 101
VW Passat B7	Diesel	77	Manual	2011	2	30, 35, 39, 40, 42, 45, 47, 49, 50, 52, 54, 55, 57, 60, 61, 64, 65, 67, 70, 71, 72, 73, 75, 78, 80, 81, 82, 85, 88, 90, 91, 94, 96, 98, 100

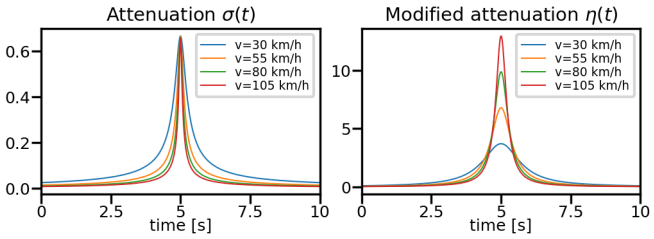


Fig. 3. *Left:* Amplitude attenuation $\boldsymbol{\sigma(t)}$ of the sound signal. *Right:* Proposed feature $\boldsymbol{\eta(t)}$ for $\boldsymbol{\alpha = v}$ and $\boldsymbol{\beta = 0.05}$. $\boldsymbol{t_{\mathrm{CPA}} = 5}$ s.

We propose to modify $\sigma(t)$ as follows:

$$\eta(t) = \frac{\alpha}{\beta v^2 (t_{\mathrm{CPA}} - t)^2 + d_{\mathrm{CPA}}^2}. \tag{2}$$

Parameter α controls the vertical extent of $\eta(t)$, whereas β affects its width. We will refer to $\eta(t)$ as *modified attenuation* (MA), and $\eta(t)$ for $v = [30, 55, 80, 105]$ km/h, $d_{\mathrm{CPA}} = 1.5$ m, $\alpha = v$ and $\beta = 0.05$ are presented in Fig. 3 (right). Clearly, α and β provide much clearer distinction between different $\eta(t)$ profiles than it is case with $\sigma(t)$.

The $\eta(t)$ feature will be predicted using the log-mel spectrogram (LMS). LMS represents the most common feature used in various acoustic pattern classification tasks [14]. It yielded excellent results in predicting clipped vehicle-to-microphone distance used for vehicle counting [15], [16].

B. Proposed method

The proposed methodology for acoustic vehicle speed estimation is illustrated in Fig. 4. From the input audio signal (top plot), LMS is calculated (second plot). Based on LMS, the proposed speed estimation feature, MA, is predicted in a supervised fashion (third plot). Each MA point is predicted using a time frame of LMS samples, as presented with the shaded area in the second plot. The pass-by instant t_{PB} is predicted by maximizing the MA profile. For speed estimation, we consider only the MA values around t_{PB} (yellow background in the third plot), the other ones contribute much less. Vehicle speed is estimated using a regression model which takes the windowed MA samples as input.

The block diagram of the proposed method is presented in Fig. 5 (top). Block *Feature extraction* outputs MA prediction file-wise. Speed estimation is carried out using the predicted MA (block *Speed estimation*).

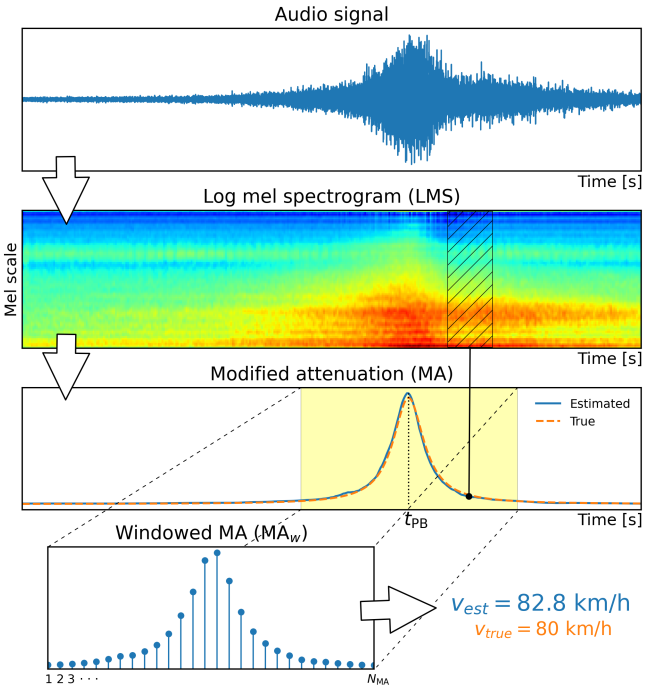


Fig. 4. Proposed methodology for acoustic vehicle speed estimation. *Top:* Original audio signal. *Second plot:* Log mel spectrogram of audio. *Third plot:* Supervised MA feature prediction based on LMS. *Bottom:* Speed is estimated using the predicted MA around its maximum.

Detailed presentation of *Feature extraction* is given in Fig.

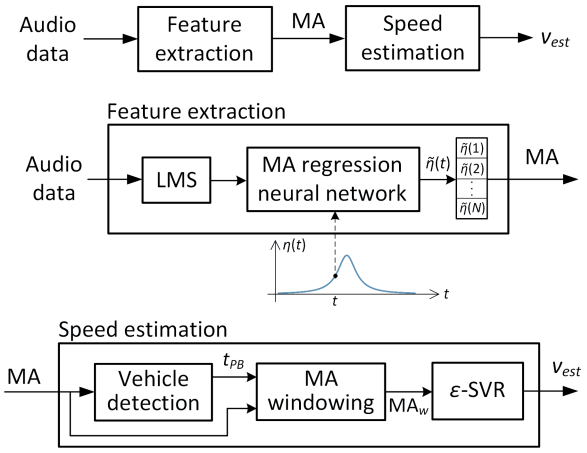


Fig. 5. *Top:* Block diagram of the proposed supervised speed estimation method. *Middle:* MA feature extraction in detail. *Bottom:* Speed estimation in detail. t_{PB} and $\mathrm{MA}_{\boldsymbol{w}}$ represent the predicted pass-by instant of vehicle and the windowed MA feature, respectively.

5 (middle). First, LMS of audio signal is calculated. Then, the MA feature is predicted based on LMS using a fully-connected neural network (NN). NN outputs a single value prediction of MA at a given time instant, $\widetilde{\eta}(t)$. To take into account time dependence between adjacent $\eta(t)$ values, $\eta(t)$ will be predicted using LMS samples from time interval $[t-Q, t+Q]$ as features (shaded area in Fig. 4, second plot).

In the *Speed estimation* block, we first detect a vehicle passing by the sensor. One detection approach is to locate peak in the sound energy [15]. However, peaks in energy can also be induced by other sound sources, such as vehicles in the nearby roads, machines in construction sites, natural sounds (wind, birds, crickets). Bearing this in mind, we adopt the MA peak magnitude to be in direct proportion with the speed, which enables vehicle detection based on maximum of the predicted MA. Setting $\alpha = v$ in (2) represents a reasonable choice in that regard, which yields the MA peak magnitude

$$\eta(t_{\mathrm{CPA}}) = \frac{v}{d_{\mathrm{CPA}}^2}. \tag{3}$$

Based on this rationale, we detect vehicle by maximizing the predicted MA profile, i.e., the pass-by instant t_{PB} will be predicted as the position of MA maximum. Setting $\alpha = v$ in (2) implies that for audio files with no vehicles passing by the sensor, the MA profile is flat and equal to zero.

As features for speed estimation, we select MA samples around t_{PB}, i.e., we carry out *MA windowing*, since the MA samples far from t_{PB} contribute much less. Selection of the window width, N_{MA}, is discussed in Section III-C.

Speed estimation is carried out using ε-support vector regression (ε-SVR) due to relatively small size of the dataset and small number of parameters to optimize. The ε-SVR parameters are C (penalty of the error term) and ε (width of the ε-insensitive zone used to fit the training data; determines the accuracy level of the approximated function) [17]. Input to the ε-SVR block is a vector of windowed MA samples (one vector per audio file), $\mathrm{MA}_w \in \mathbb{R}^{N_{\mathrm{MA}}}$, and the output is speed estimation v_{est}.

C. Implementation details

1) Modified attenuation: In (2), we set $\alpha = v$ (discussed in Section III-B), $\beta = 0.05$ and $d_{\mathrm{CPA}} = 1.5$ m. The selected β value gave the most accurate speed estimation over a grid of β values. The selected d_{CPA} is close to the actual distance between the vehicle and the sensor in the experiment. LMS is based on the short-time Fourier transform (STFT) of the input signal. For the STFT calculation, we use the Hamming window with $N_w = 4096$ samples (≈ 0.093 s) and the hop length of $N_h = 0.27N_w = 1105$ samples (≈ 0.025 s). With 10-second audio files sampled at 44100 Hz, this setup gives the time-length of STFT of 400 time frames. In the LMS calculation, $N_{mel} = 40$ mel bands are used within the frequency band $[0, 16$ kHz$]$. For the regression of $\eta(t)$, as input to NN we take LMS at instant t and $Q = 12$ preceding and following instants with a stride of 3. The input space dimensionality is hence $M = (2Q+1)N_{mel} = 1000$. NN has five layers, with 1000–200–50–10–1 neurons per layer, respectively. This configuration is nearly minimal in the sense that further increasing the number of layers and neurons per layer would not yield any significant improvement in the MA regression accuracy tested on the validation data. Mean squared error loss is used, ReLU activation (last layer uses linear activation), $L2$ kernel regularization with factor 10^{-3}, 200 training epochs.

2) Speed estimation: The optimal window length and ε-SVR parameters are $N_{\mathrm{MA}} = 73$, $C = 10$ and $\varepsilon = 0.1$, obtained via a three-dimensional grid search. For training and validating the ε-SVR model, same train-validation split is used as in the MA regression.

Method implementation in Python is available for download at `http://slobodan.ucg.ac.me/science/vse/`.

D. Evaluation

We carry out 10-fold cross-validation. In one round of cross-validation, one fold (vehicle) is retained for testing, whereas the remaining nine folds are used for training and validating the model (train-validation split is described in Section II-B). The cross-validation process is repeated 20 times and the averaged results are presented in Section IV.

IV. Experimental Results

We evaluate the speed estimation performance using root-mean-square error (RMSE) of speed estimation

$$\mathrm{RMSE} = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(v_l^{est} - v_l^{true})^2}, \tag{4}$$

where v_l^{est} and v_l^{true} represent the values of estimated and true speed of the l-th measurement (audio file), respectively, whereas L represents the number of measurements.

For the second evaluation metric, we will discretize the considered speed interval $[30, 105]$ km/h with a step of 10 km/h, starting from 25 km/h (first interval $[25, 35)$, second $[35, 45)$ and so on). Then, the vehicle sound is classified into these speed intervals (speed classes). The second metric will be the accuracy of speed classification expressed as probability of predicting a speed class that is Δ classes away from the true

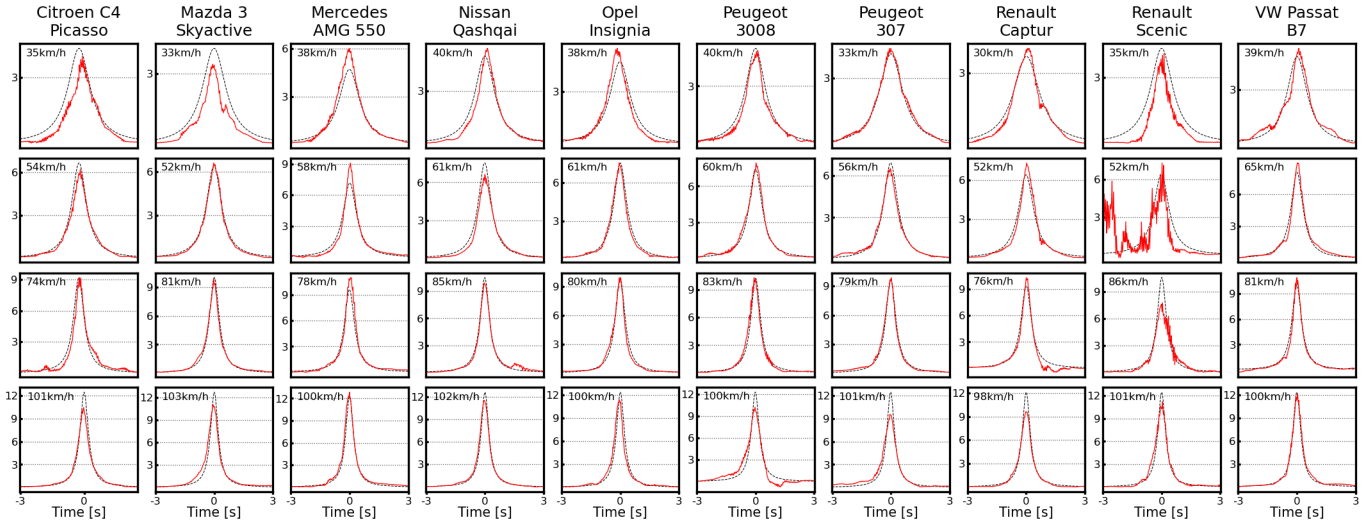


Fig. 6. Predictions (in red) of the proposed MA feature (test data, one run). Four speeds per vehicle are presented (speed given top left).

class ($\Delta = 0$ class prediction is correct, $\Delta = \pm 1$ predicted and true class are adjacent ones, and so on).

In Fig. 6, we present one run of MA predictions of the test data. Four speeds per vehicle are presented, with speed given top left. The MA plots are centered with respect to the predicted pass-by instant t_{PB}. The lowest MA prediction performance is obtained with Renault Scenic. One Renault Scenic recording session took place on a windy day, which resulted in a strong wind corrupting the sound of the vehicle in around a half of the recordings. Moderate to strong wind is present in some of Citroen, Mercedes, Peugeot 3008 and Passat recordings.

Let us first consider the vehicle detection performance. To that end, we calculate detection error as difference of positions of the true and predicted MA maxima of the test data. Histogram of detection error (all 20 runs included) is presented in Fig. 7 (top). Detection error can be modeled as a normal random variable, with mean and standard deviation presented top right. The histogram also shows that the proposed method is able to accurately detect the pass-by instant of vehicle, with absolute detection error less than 0.2 s. In Fig. 7 (bottom), we compare the predicted MA maxima values of the test data when audio files i) contain vehicles (blue histogram) and ii) do not contain vehicles (orange histogram) passing by the sensor. To that end, we introduce additional 35 no-vehicle sound files used only for vehicle detection testing. The additional files were tested with NN regression models of all vehicles in all 20 runs, 200 models in total. Histograms of the vehicle and no-vehicle cases are separated by a narrow green rectangle in Fig. 7 (bottom), i.e., all no-vehicle cases are to the left of the rectangle, all vehicle cases to the right. The MA magnitude threshold for vehicle detection should be set within the rectangle.

RMSEs of speed estimation are presented in Table II, per vehicle and average (bottom). The result of Renault Scenic is notably worse than of the other vehicles, which is due to a strong wind corrupting the sound, as noted above. If we exclude Renault Scenic in testing, the average RMSE will be

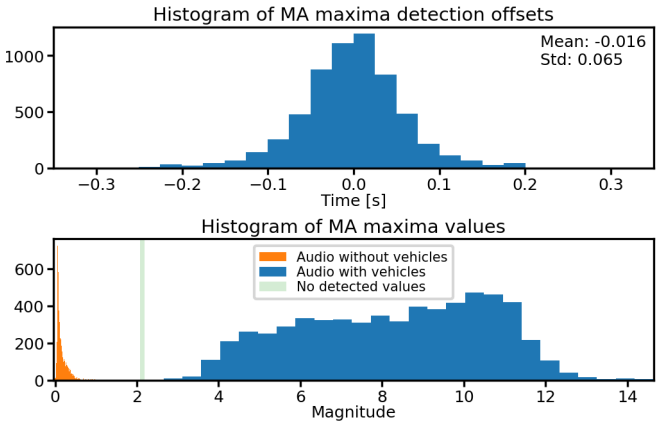


Fig. 7. *Top:* Histogram of MA maxima detection offsets. *Bottom:* Histogram of MA maxima values. Green rectangle separates no-vehicle (orange) and vehicle (blue) cases.

6.95 km/h. On the other hand, Opel Insignia speed estimations are exceptionally accurate.

Figure 8 represents 95% confidence intervals for the mean of speed estimation. Renault Scenic is characterized by the most significant deviation from the true speeds, as opposed to Opel Insignia which follows the true speeds faithfully. A general trend that can be observed in Fig. 8 is that speed estimation at low and medium speeds is more accurate than at higher speeds. More precisely, the proposed method tends to underestimate higher speeds, which is evident from the plots of Citroen, Mazda, Peugeot 3008, Peugeot 307, and Renault Captur. A slight underestimation is present also with Opel. The underestimation issue will be discussed shortly.

Table III presents the classification accuracies when speed estimation is formulated as a classification problem. Column $\Delta = 0$ in Table III corresponds to correct class prediction and $|\Delta| \leq 1$ corresponds to a misclassification of maximum one class. Renault Scenic has the lowest $|\Delta| \leq 1$ probability, whereas Opel Insignia outperforms all other vehicles, with a near-perfect accuracy of 99.8%. Nissan and Mercedes are also very accurate in terms of the $|\Delta| \leq 1$ probability, however

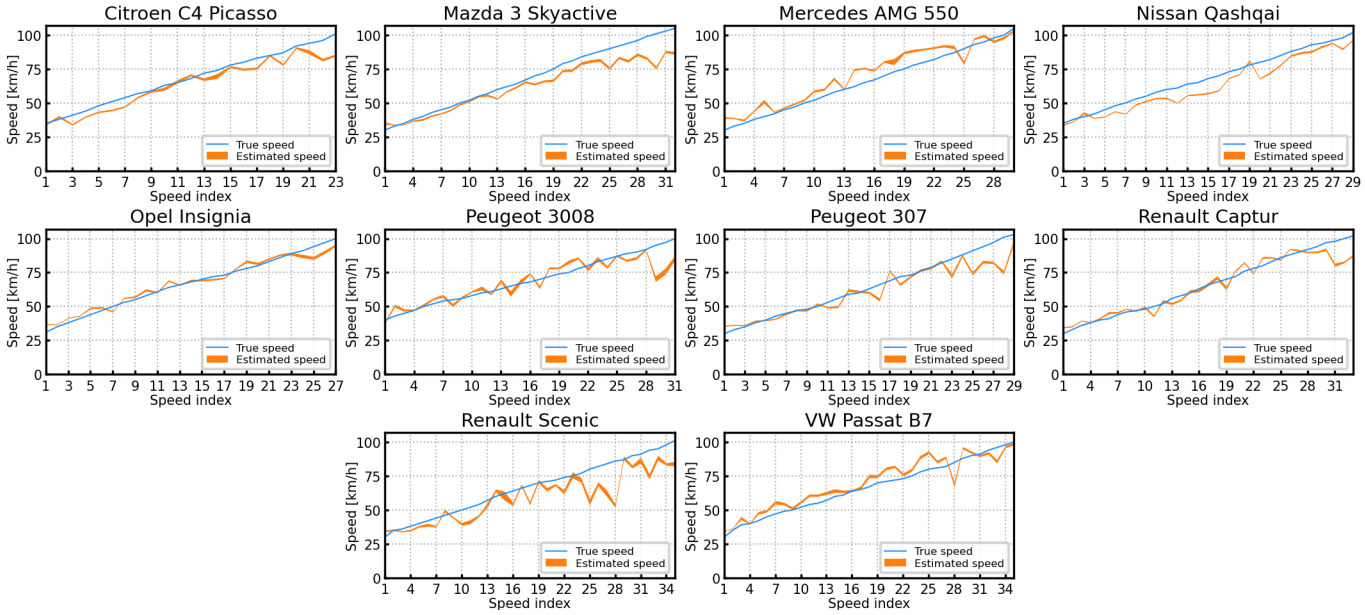


Fig. 8. 95% confidence intervals for the mean of speed estimation. Speed index on the horizontal axis represents the index of the recorded speed; speeds are sorted into ascending order, as listed in Table I.

TABLE II
RMSE OF SPEED ESTIMATION

Vehicle	RMSE [km/h]
Citroen C4 Picasso	6.48
Mazda 3 Skyactive	8.86
Mercedes AMG 550	7.56
Nissan Qashqai	6.53
Opel Insignia	3.97
Peugeot 3008	8.02
Peugeot 307	8.12
Renault Captur	6.22
Renault Scenic	11.34
VW Passat B7	6.76
Average	7.39

their $\Delta = 0$ probabilities are much less than Opel's 74.3%.

TABLE III
PROBABILITY OF PREDICTING CLASS THAT IS Δ CLASSES AWAY FROM THE TRUE CLASS

Vehicle	$\Delta = 0$	$\lvert\Delta\rvert = 1$	$\lvert\Delta\rvert = 2$	$\lvert\Delta\rvert > 2$	$\lvert\Delta\rvert \leq 1$
Citroen C4 Picasso	61.3%	32.2%	6.5%	0.0%	93.5%
Mazda 3 Skyactive	41.6%	48.6%	8.8%	1.1%	90.2%
Mercedes AMG 550	48.7%	49.5%	1.8%	0.0%	98.2%
Nissan Qashqai	40.0%	59.5%	0.5%	0.0%	99.5%
Opel Insignia	74.3%	25.6%	0.2%	0.0%	99.8%
Peugeot 3008	48.5%	42.9%	4.8%	3.7%	91.5%
Peugeot 307	61.9%	24.8%	11.4%	1.9%	86.7%
Renault Captur	57.9%	36.2%	5.9%	0.0%	94.1%
Renault Scenic	42.9%	41.6%	11.1%	4.4%	84.4%
VW Passat B7	55.4%	40.9%	3.0%	0.7%	96.3%
Average	53.2%	40.2%	5.4%	1.2%	93.4%

Limitation of the proposed MA feature is that it is symmetric with respect to the pass-by instant. This symmetry, however, is not present in the actual amplitude attenuation of the pass-by sound, especially at higher speeds. Namely, as the speed increases, noise due to air flow generated by the boundary layer of the vehicle, perceived as a whoosh sound, becomes an important factor in the overall loudness of vehicle [5]. This noise is prominent immediately after the vehicle passes by the microphone. Figure 4 (top) indicates asymmetry in sound attenuation before and after the pass-by instant. This asymmetry is even more pronounced at higher speeds. Not taking into account this phenomenon in design of the MA feature probably represents a reason of less accurate estimation at higher speeds. Other reasons could be suboptimal analytical form of the MA feature (2) and suboptimal selection of the coefficients α and β in (2).

V. CONCLUSIONS

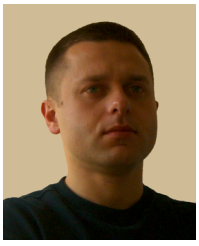
We proposed a method for vehicle speed estimation based on the sound which a vehicle produces while passing by the sensor. Our speed estimation uses a novel speed-dependent feature predicted from the input audio. The method is trained and tested on the collected dataset of audio-video recordings of vehicles passing by the sensor. The method accurately detects a vehicle and estimate its speed with an average error of 7.39 km/h. When formulated as a classification problem, i.e., when the speed is discretized into 10 km/h intervals, the achieved accuracy is 53.2% for correct interval prediction and 93.4% when misclassification of one interval is allowed.

The proposed method tends to underestimate higher speeds. Therefore, the future research will aim to improve the estimation accuracy at higher speeds by modifying the proposed feature and/or by introducing additional features. In addition, data augmentation techniques will be considered to improve the estimation accuracy, as well as extending the dataset by introducing new vehicles.

REFERENCES

[1] Myounggyu Won, "Intelligent traffic monitoring systems for vehicle classification: A survey," *IEEE Access*, vol. 8, pp. 73340–73358, 2020.

[2] Cecilia Wilson, Charlene Willis, Joan K Hendrikz, Robyne Le Brocque, and Nicholas Bellamy, "Speed cameras for the prevention of road traffic injuries and deaths," *Cochrane database of systematic reviews*, , no. 11, 2010.
[3] BG Quinn, "Doppler speed and range estimation using frequency and amplitude estimates," *The Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2560–2566, 1995.
[4] Christophe Couvreur and Yoram Bresler, "Doppler-based motion estimation for wide-band sources from single passive sensor measurements," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1997, vol. 5, pp. 3537–3540.
[5] Volkan Cevher, Rama Chellappa, and James H McClellan, "Vehicle speed estimation using acoustic wave patterns," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 30–47, 2008.
[6] Shubhranshu Barnwal, Rohit Barnwal, Rajesh Hegde, Rita Singh, and Bhiksha Raj, "Doppler based speed estimation of vehicles using passive sensor," in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 2013, pp. 1–4.
[7] Elżbieta Kubera, Alicja Wieczorkowska, Andrzej Kuranc, and Tomasz Słowik, "Discovering speed changes of vehicles from audio data," *Sensors*, vol. 19, no. 14, pp. 3067, 2019.
[8] Hendrik Vincent Koops and Franz Franchetti, "An ensemble technique for estimating vehicle speed and gear position from acoustic data," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 2015, pp. 422–426.
[9] Jader Giraldo-Guzmán, Andrés Guillermo Marrugo, and Sonia H Contreras-Ortiz, "Vehicle speed estimation using audio features and neural networks," in *2016 IEEE ANDESCON*. IEEE, 2016, pp. 1–4.
[10] Hüseyin Göksu, "Vehicle speed measurement by on-board acoustic signal processing," *Measurement and Control*, vol. 51, no. 5-6, pp. 138–149, 2018.
[11] Roberto López-Valcarce, Carlos Mosquera, and Fernando Pérez-González, "Estimation of road vehicle speed using two omnidirectional microphones: A maximum likelihood approach," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 8, pp. 929146, 2004.
[12] Kam W Lo and Brian G Ferguson, "Broadband passive acoustic technique for target motion parameter estimation," *IEEE Transactions on aerospace and electronic systems*, vol. 36, no. 1, pp. 163–175, 2000.
[13] Patrick Marmaroli, Jean-Marc Odobez, Xavier Falourd, and Herve Lissek, "Pass-by noise acoustic sensing for estimating speed and wheelbase length of two-axle vehicles," in *Proceedings of Meetings on Acoustics ICA2013*. Acoustical Society of America, 2013, vol. 19.
[14] Romain Serizel, Victor Bisot, Slim Essid, and Gaël Richard, "Acoustic features for environmental sound analysis," in *Computational Analysis of Sound Scenes and Events*, pp. 71–101. Springer, 2018.
[15] Slobodan Djukanović, Jiří Matas, and Tuomas Virtanen, "Robust audio-based vehicle counting in low-to-moderate traffic flow," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1608–1614.
[16] Slobodan Djukanović, Yash Patel, Jiří Matas, and Tuomas Virtanen, "Neural network-based acoustic vehicle counting," in *29th European Signal Processing Conference (EUSIPCO 2021)*, 2021.
[17] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.



Slobodan Djukanović received the M.Sc. and Ph.D. degrees in electrical engineering from the University of Montenegro, Podgorica, in 2004 and 2008, respectively. He is currently a Full Professor at the Faculty of Electrical Engineering, University of Montenegro. During 2008/2009, he was in Grenoble, France, where he finished his post-doctoral studies at GIPSA-lab, CNRS. During 2019/2020, he was full-time employed as a researcher at Czech Technical University, Faculty of Electrical Engineering, Department of Cybernetics on project "Audio-visual object classification and sound event recognition by unsupervised co-training" project. His research concerns digital signal processing, machine learning, audio-based traffic monitoring. More information at http://www.tfsa.ac.me/slobodan.html.

Jiří Matas received the Ph.D. degree from the University of Surrey, Guildford, U.K., in 1995. He is a Professor with Center for Machine Perception, Czech Technical University, Prague, Czech Republic. He has published more than 200 articles in refereed journals and conferences. His publications have approximately 53000 citations in Google Scholar and 17,000 in the Web of Science. His H-index is 85 (GS) and 51 (WoS), respectively. His research interests include visual tracking, object recognition, image matching and retrieval, sequential pattern recognition, and RANSAC-type optimization methods. He is on the editorial board of the IJCV and was an Associate Editor-in-Chief of the IEEE Transactions on Pattern Analysis and Machine Intelligence.

Tuomas Virtanen is Professor at Tampere University, Finland, where he is leading the Audio Research Group. He received the M.Sc. and Doctor of Science degrees in information technology from Tampere University of Technology in 2001 and 2006, respectively. He has also been working as a research associate at Cambridge University Engineering Department, UK. He is known for his pioneering work on single-channel sound source separation using non-negative matrix factorization based techniques, and their application to noise-robust speech recognition and music content analysis. Recently he has done significant contributions to sound event detection in everyday environments. In addition to the above topics, his research interests include content analysis of audio signals in general and machine learning. He has authored more than 200 scientific publications on the above topics, which have been cited more than 13000 times. He has received the IEEE Signal Processing Society 2012 best paper award for his article "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria" as well as three other best paper awards. He is an IEEE Fellow, member of the Audio and Acoustic Signal Processing Technical Committee of IEEE Signal Processing Society, and recipient of the ERC 2014 Starting Grant.