Acoustic Traffic Monitoring Based on Deep Neural Network Trained by Stereo-Recorded Sound and Sensor Data

Tomohiro Takahashi, Yuma Kinoshita, Yukoh Wakabayashi, and Nobutaka Ono
Tokyo Metropolitan University, Tokyo, Japan
Jun Honda, Seishi Fukuma, Aoi Kitamori and Hiroshi Nakagawa
NEXCO-EAST ENGINEERING, Tokyo, Japan

***Abstract*—In this study, we present a machine-learning-based acoustic traffic monitoring aiming to realize a relatively low-cost system compared with existing traffic sensors. Since vehicles are moving fast and the sound from different vehicles may overlap, the relationship between acoustic signals and traffic conditions, such as the number of vehicles passing or speed, is complicated. Then, the machine-learning approach is attractive. For this purpose, collecting a sufficient amount of data to train, for example, deep neural networks (DNNs), is crucial. In this study, we built a 48-hours dataset using stereo microphones and sensors already installed on highways to label traffic conditions automatically. We used ConvMixer, one of the recently well-used convolutional neural network (CNN) architectures, to estimate four traffic conditions, i.e., the total number of vehicles passing, the number of large vehicles passing, speed, and time occupancy. In experiments, we compare the acoustic features used as input to the DNN, compare our method with conventional methods, and apply our method to traffic flow discrimination.**

***Index Terms*—traffic monitoring, vehicle detection, convolutional neural network, acoustic sensing, microphone array**

I. Introduction

Measuring road traffic information, such as traffic volume, speed, and time occupancy, is essential to understanding traffic conditions in real time and providing information to road traffic control and users. For this reason, expressway companies use traffic detectors to measure traffic information [1], [2].

The most common traffic detector on Japanese intercity highways is a loop-coil-based one, which electro-magnetically detects the passage of vehicles through loop coils embedded in the road pavement [3]. Although they have high observation accuracy, their installation and operation costs are high. Therefore, the installation of loop coils is limited, only at one point per interchange in rural areas with low traffic volume, for example. The realization of a low-cost traffic detector would be desired for highly dense traffic monitoring. Aiming to this, we investigate acoustic traffic monitoring systems in this study since it is much easier to install microphones on the roadside.

Several methods, such as rule-based vehicle detection [4]–[8] and machine learning [9]–[12], have been proposed for traffic monitoring using acoustic information. However, rule-based methods have yet to adequately model the complex relationships between acoustic features and traffic conditions on real roads. In addition, machine learning methods should consider how to collect sufficient amounts of labeled training data and input features.

In this study, running-vehicle sound signals were recorded near traffic detectors installed on an actual highway to construct an automatically labeled traffic condition dataset. Using this dataset, we used a deep neural network (DNN) to estimate four traffic conditions, including the number of vehicles passing and their speeds. To evaluate the performance of our method, in this study, we investigated how the processing of running-vehicle sound signals can be helpful for analysis by conducting experiments in which we compared acoustic features as input to the DNN, compared our method with conventional methods, and applied our method to traffic flow discrimination. The best performance was obtained when the power spectrogram and phase difference were input to the DNN as acoustic features. Our proposed method showed higher traffic condition estimation accuracy than conventional methods. In addition, good accuracy was confirmed in discrimination of congested and free flow.

This work was supported by JSPS KAKENHI Grant Number JP20H00613.

II. Related Work

Conventional traffic monitoring methods using acoustic information are classified into rule-based and machine-learning-based methods. These methods are briefly summarized below.

A. Rule-based Methods

Sobreira-Seoane *et al.* proposed a rule-based method for vehicle detection [4]. This method is based on the short-time power variation of sound detected by a single-channel microphone. Such a method using single-channel signals is advantageous in terms of microphone installation cost, but it cannot utilize the spatial information of sound.

There are various methods using multichannel signals: vehicle detection by analyzing the intensity of the sound recorded by a microphone array [5]; speed estimation [6]; and the estimation of the number of vehicles using the energy ratio of the peaks between channels [7]. The method using the time difference of arrival (TDOA) between channels has also been studied by Ishida *et al.* [8] This method counts the number

Fig. 1. Recording scenery. Traffic detector is embedded in road pavement in direction of travel, and is not shown in picture.

of vehicles, considering the fact that TDOA monotonically increases (decreases) as vehicles approach and forms an S-shaped curve. However, the vehicles cannot be accurately counted by this method when TDOA complexly changes, such as in the case when multiple vehicles are traveling or when the sounds of vehicles in opposite lanes are recorded, making it challenging to apply this method to roads with high traffic volume or numerous lanes.

B. Machine-Learning-Based Methods

Machine-learning-based traffic monitoring with acoustic information has been widely studied. For example, the k-nearest neighbor method and neural networks are applied to acoustic signals to detect vehicles and classify vehicle types, respectively [9], [10]. Tyagi *et al.* proposed a method of classifying traffic density states (i.e., congested, medium, and free flow) using support vector machines [11]. In addition, a method involving vehicle detection and speed estimation using a four-layer fully connected DNN has also been proposed [12]. This method uses the power and TDOA of running-vehicle sound signals on highways as input features.

For machine learning methods, collecting datasets for training a model is challenging. For example, Shinohara *et al.* [12] manually annotated the time and average speed of vehicles passing on the basis of road images captured by two video cameras next to microphones for recording vehicle sounds. As a result, the total size of the constructed dataset was small (7 h), and there were also annotation errors due to the calculation of the average speed from the videos.

Considering these existing related studies, this study used the machine learning approach. To collect sufficient training data, we constructed a dataset with automated traffic condition labeling instead of manual annotation by recording traffic sounds near traffic detectors installed on actual highways. In addition, as described below, a recently proposed network structure, called ConvMixer [13], was introduced to machine learning for traffic monitoring.

III. Dataset Construction

To train a DNN for traffic monitoring based on sound, a new dataset consisting of 48 h of acoustic signals recorded on a highway and traffic conditions measured by traffic detectors installed near the recording points was constructed.

The acoustic signals were recorded near the 6.7-km post on Yokohama Shindo in Yokohama City, Kanagawa Prefecture, Japan, for 4 h from 6:00 am to 10:00 am on 12 weekdays from October 11 to 26, 2021. The road at the measurement point consisted of two inbound lanes: a travel lane and an overtaking lane[1]. The acoustic signals were recorded by installing a Panasonic RR-XS470 IC recorder at a distance of approximately 170 m from the traffic detector[2] (see Fig. 1). The IC recorder was installed such that the two microphones mounted on it were parallel to the direction of the road. Since it was not known whether the high-pass filter of the IC recorder could remove the effect of wind noise, two units were prepared, one with the high-pass filter turned on and the other with the filter turned off, and the acoustic signals were recorded using these two units. The sampling frequency Fs was set to 44.1 kHz.

Traffic conditions were measured by traffic detectors installed at approximately 170 m from the IC recorder. Traffic conditions for each lanes were obtained every one minute from the traffic detectors: the total number of vehicles passing, the number of large vehicles passing, the speed of vehicles, and the time occupancy (OCC)[3] in the travel and overtaking lanes, with the standard time information. The [1st, 3rd] quartiles for each traffic condition for all measured data are [43, 55] (unit), [1, 4] (unit), [42.6, 73.4] (km/h), and [9, 18] (%), respectively. We synchronized the time of the traffic detector and the recorded sounds by playing a special sound at a specific time such as 5:59 am.

IV. Network Architecture and Features

For traffic monitoring using acoustic information, in this study, we used a convolutional neural network (CNN), called ConvMixer. Specifically, the acoustic features of running-vehicle sound were input to the network to estimate four traffic conditions (total number of vehicles passing, number of large vehicles passing, speed, and OCC). In training, the traffic conditions observed by the traffic detectors were used as a teacher to calculate the loss from the CNN estimates. In the following sections, we refer to this method as the proposed method and describe its main components.

A. Input Features

The proposed method uses stereo running-vehicle sound as the input acoustic signal, as in previous studies [8], [12], because it is considered that spatial sound information is effective in traffic monitoring using acoustic information.

[1]Totally, there are four lanes but we did not focused on the two outbound lanes. The traffic detector measures conditions on the two inbound lanes separately.

[2]The IC recorder should have been placed directly above the traffic detector. However, there was no shoulder to install; thus, the recording was conducted this way. The time difference between the traffic detector and the stereo IC recorder is 9.24 s when the vehicle speed is 66.2 km/h. This time difference is not considered in this study since it is much smaller than one minute.

[3]Ratio of the time a vehicle steps on the loop coil of the traffic detector to the observed time

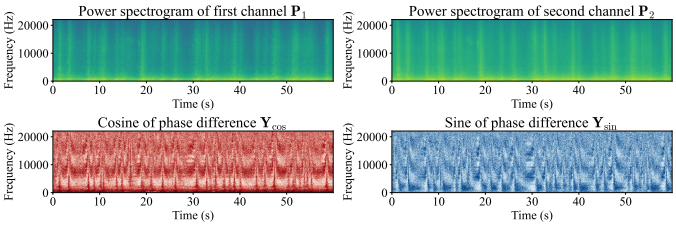

Fig. 2. 2-channel power spectrograms and cosine and sine of phase difference

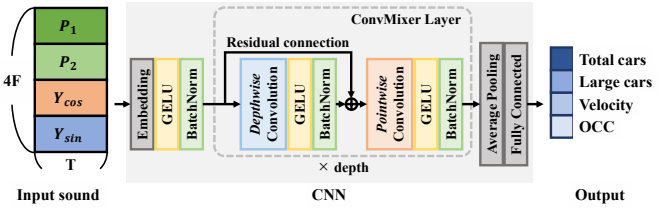

Fig. 3. Network architecture used in our experiments

Here, the input acoustic signal is divided into 1-min segments since the traffic detector observations used by a teacher are obtained every minute. From the comparison of several different acoustic features as input in Section V-A, in this section, we describe the case where the best-performing input features are the 2-channel power spectrograms $\mathbf{P}_1$ and $\mathbf{P}_2$ and the cosine and sine of the phase differences $\mathbf{Y}_{\cos}$ and $\mathbf{Y}_{\sin}$ (see Fig. 2) calculated using the following equations:

$$\mathbf{X}_i = \mathrm{STFT}(\mathbf{x}_i), \tag{1}$$

$$P_i(k,l) = |X_i(k,l)|^2, \tag{2}$$

$$\Delta\phi(k,l) = \arg(X_1(k,l)/X_2(k,l)), \tag{3}$$

$$Y_{\cos}(k,l) = \cos(\Delta\phi(k,l)), \tag{4}$$

$$Y_{\sin}(k,l) = \sin(\Delta\phi(k,l)). \tag{5}$$

Here, $\mathbf{x}_i$ is the input acoustic signal of the ith channel ($i = 1, 2$), and $\mathrm{STFT} : \mathbb{C}^N \to \mathbb{C}^{F\times T}$ in (1) represents the short-time Fourier transform that transforms an N sample time signal into a spectrogram with a frequency bin number $F\times$ time frame number T. k and l are indices in the frequency and time directions, respectively, and $X_i(k,l)$ denotes the (k,l) element of $\mathbf{X}_i$ (the same notation applies to other matrices). As for the phase difference in (3), the effect of its periodicity on the input features is ignored by computing cos and sin as in (4) and (5). The arrays $(\mathbf{P}_1^\top, \mathbf{P}_2^\top, \mathbf{Y}_{\cos}^\top, \mathbf{Y}_{\sin}^\top)^\top$, which are concatenated in the frequency index direction, are input to the CNN.

B. Network Architecture

In traffic monitoring, in addition to mixing features for each time frame, it is helpful to mix features before and after the time frame of interest; therefore, we use ConvMixer as the CNN architecture in this study [13], [14]. ConvMixer has a repeating network structure consisting of a Pointwise Convolution, which mixes the features for each time frame using a Fully Connected layer, and a Depthwise Convolution, which mixes the features by convolving in the direction of the time frame (see Fig. 3). Patch embedding is applied to input features by handling short time frames in the input as patches. The number of embeddings is set to T. The dimension of feature vectors in the ConvMixer Layer is 529, and the kernel size of the Depthwise Convolution is 5. The number of iterations of the ConvMixer Layer is set to 5. These parameters are determined experimentally. The output of the ConvMixer Layer is finally aggregated in the time-frame direction by the Average Pooling layer and then transformed into a four-dimensional vector representing the traffic conditions by the Fully Connected layer.

C. Output Variables

The output variables of the CNN are the four traffic conditions that can be measured by the traffic detectors used in the dataset described in Section III: the total number of vehicles passing, the number of large vehicles passing, speed, and OCC. These traffic conditions were indivisually measured for each lane, and they were aggregated. The same traffic conditions as those for the traffic detectors are obtained by sound traffic monitoring, and multi-task learning, which estimates multiple traffic conditions, is used to improve accuracy. Since the range of values is different for each traffic condition, normalization was applied during training so that the mean is 0 and the variance is 1 for each traffic condition.

V. Experiments

To evaluate the proposed method, we investigated the change in estimation performance depending on the acoustic features an input to the DNN. Experiments were also conducted to compare the estimation performance of the proposed method with that of conventional methods and to apply the proposed method to traffic flow discrimination using the estimator of the proposed method.

A. Comparison of Input Features

1) Experimental Conditions: In this experiment, data from two days, October 12 and 21, of the dataset described in Section III were used for evaluation. Four traffic conditions (total number of vehicles passing, number of large vehicles passing, speed, and OCC) were estimated every minute on these two days. Their accuracy was evaluated in terms of root mean squared error (RMSE) with the observed values from traffic detectors.

For training the CNN, we used seven days of data for training and two days for validation out of nine days of data, excluding the two days of data used for evaluation. The loss function for training was set to the mean absolute error, and Adam [15] was used as the optimization function with an initial learning rate of 0.0005 for 100 epochs.

To investigate suitable feature extraction methods for traffic monitoring, we trained the CNN described in Section IV by combining several feature extraction methods and compared the input features in terms of the RMSE of the estimated values. The following three amplitude-related features were considered:

TABLE I
COMPARISON OF INPUT FEATURES
(MEAN RMSE SCORE ± STD. FOR 5 TRIALS)

	Total number of vehicles (unit)	Number of large vehicles (unit)	Speed (km/h)	OCC (%)
Pow	6.39±0.32	1.88±0.05	6.66±0.41	2.83±0.05
TF-Pow	5.04±0.25	1.63±0.03	6.15±0.34	2.6±0.05
LogTF-Pow	5.8±0.42	1.64±0.03	6.25±0.42	2.77±0.09
Pow+TDOA	5.93±0.64	1.93±0.06	6.28±0.59	2.58±0.11
TF-Pow+TDOA	5.24±0.27	1.63±0.05	5.9±0.4	2.65±0.19
TF-Pow+WhTDOA	5.2±0.14	1.62±0.03	6.24±0.63	2.65±0.2
TF-Pow+PhaseDiff	**4.67**±0.18	**1.52**±0.03	**5.48**±0.3	**2.54**±0.07
Time domain	6.73±0.08	1.79±0.05	7.13±0.2	2.99±0.08

Pow: $\mathbf{f}_{\mathrm{pow}} \in \mathbb{R}^{1\times T}$ whose lth element is the power averaged in frequency and channel direction as

$$f_{\mathrm{pow}}(l) = \frac{1}{2}\sum_{i=1}^{2}\sum_{k=0}^{F/2} P_i(k,l) \tag{6}$$

TF-Pow: $\mathbf{F}_{\mathrm{tfp}} = \mathbf{P}_i \in \mathbb{R}^{F\times T}$, where the (k,l) element is $P_i(k,l)$ shown in (2)

LogTF-Pow: $\mathbf{F}_{\mathrm{ltfp}} \in \mathbb{R}^{F\times T}$, which is the common logarithm of $\mathbf{P}_i$

In addition, the three phase-related features below were considered.

TDOA: $\mathbf{f}_{\mathrm{tdoa}} \in \mathbb{R}^{1\times T}$ whose lth element is the time difference of arrival of sound obtained by cross-correlation from two-channel signals as

$$f_{\mathrm{tdoa}}(l) = \arg\max_n \psi_l[n], \tag{7}$$

$$\psi_l = \mathrm{IDFT}(\Psi(\cdot,l)), \tag{8}$$

$$\Psi(k,l) = X_1^*(k,l)\ X_2(k,l) \tag{9}$$

WhTDOA: $\mathbf{f}_{\mathrm{wtdoa}} \in \mathbb{R}^{1\times T}$ whose lth element is TDOA of sound obtained by cross-correlation phase transformation from two-channel signals as

$$f_{\mathrm{wtdoa}}(l) = \arg\max_n \psi'_l[n], \tag{10}$$

$$\psi'_l = \mathrm{IDFT}(\Psi'(\cdot,l)), \tag{11}$$

$$\Psi'(k,l) = \frac{X_1^*(k,l)\ X_2(k,l)}{|X_1^*(k,l)\ X_2(k,l)|} \tag{12}$$

PhaseDiff: $\mathbf{F}_{\mathrm{pd}} = (\mathbf{Y}_{\cos}^\top, \mathbf{Y}_{\sin}^\top)^\top \in \mathbb{R}^{2F\times T}$

STFT was performed in half overlap with a frame length N of 4096 points (approximately 93 ms) for a 1-min signal. When signal length $L = 60\,\mathrm{s}$, $F = \lfloor N/2 \rfloor + 1 = 2049$ and $T = \lceil L \cdot Fs/(N/2) \rceil = 1292$. After finding the $n \in [0, 2N-2)$ of the discrete that maximizes ψ, the TDOA per subsample was calculated by quadratic interpolation [16]. $\Psi(\cdot,l)$ is the l-th column vector of Ψ (the same notation applies to Ψ').

2) Experimental Results: Table I shows the results of RMSE evaluation of the estimation performance of each feature extraction method when applied to each traffic condition. A comparison of amplitude feature extraction methods confirms that **TF-Pow** performs better than the other two feature extraction methods. The results suggest that it is adequate to input power spectrograms without averaging or logarithmizing the values for each channel and time–frequency domain. When **TF-Pow** was combined with a feature extraction method for phase, there was no significant difference in estimation performance when **TDOA** and **WhTDOA** were input compared with the case when only **TF-Pow** was input. On the other hand, when **TF-Pow** was combined with **PhaseDiff**, the estimation performance was improved. The results indicate that phase information is adequate for traffic conditions estimation, and it is desirable to input the phase information of each time–frequency domain while retaining it. The estimation performance was poorest when the time domain waveforms were input as they were, confirming that conversion to the time–frequency domain is an effective means of feature extraction.

TABLE II
COMPARISON WITH CONVENTIONAL METHODS
(MEAN RMSE SCORE ± STD. FOR 5 TRIALS)

	Total number of vehicles (unit)	Number of large vehicles (unit)	Speed (km/h)	OCC (%)
Ishida *et al.* [8]	21.6±0	—	—	—
Shinohara *et al.* [12]	7.04±0.32	—	6.88±0.07	—
Proposed	**4.67**±0.18	1.52±0.03	**5.48**±0.3	2.54±0.07

B. Comparison to Related Work

1) Experimental Conditions: In this experiment, as in the experimental conditions in Section V-A, data from two days, October 12 and 21, were used for evaluation. The results of the minute-by-minute estimation of traffic conditions were evaluated in terms of RMSE with the observed values from traffic detectors. As input features of the proposed method, we used a combination of **TF-Pow** and **PhaseDiff**, which was confirmed to be the most effective in the experiments described in Section V-A, and the training conditions of the CNN were also the same as those in Section V-A.

In the comparison studies, we used the rule-based method of Ishida *et al.* [8] and the DNN-based method of Shinohara *et al.* [12] Ishida *et al.*'s method requires the parameters for the three subcurves detection of the TDOA curve. In this experiment, we experimentally determined them to obtain the highest estimation accuracy for the evaluation data.

Shinohara *et al.* [12] used a dataset labeled every 5 s. They used power and TDOA every 5 s as input features to estimate the total number and speed of vehicles passing. On the other hand, in the dataset used in this experiment, labels are assigned every minute. Therefore, in learning and evaluating Shinohara *et al.*'s method [12], we calculated the 1-min output using the 5-s estimates and compared it to the correct label. Specifically, the 1-min output number of vehicles passing was calculated by adding up the 5-s estimates, and the speed was the arithmetic mean of the 5-s estimates. Shinohara *et al.* used the information on oncoming traffic. Still, the dataset used in their experiment did not include information on oncoming lane labels; therefore, we did not use this information. Other training conditions are the same as those in the proposed method.

2) Experimental Results: Table II shows the results of the RMSE evaluation of the estimation performance of each

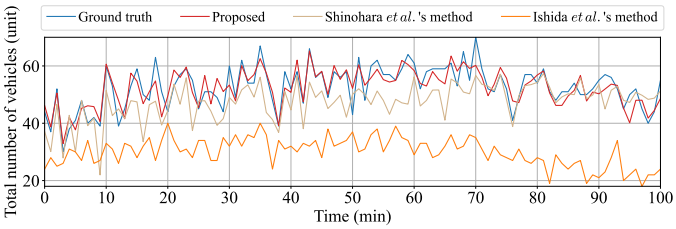

Fig. 4. Temporal change in estimated total number of vehicles

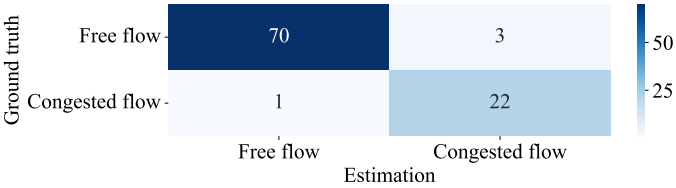

Fig. 5. Confusion matrix of traffic flow discrimination

method for each traffic condition. From Table II, it can be seen that the proposed method can estimate the total number and speed of vehicles passing with a higher accuracy than the other methods. In addition, as can be seen from Fig. 4, there is a significant difference in estimation performance between the rule-based and machine learning methods.

C. Discrimination of Traffic Flow

1) Experimental Conditions: We applied the best-performing estimator trained under the conditions specified in Section V-B to discriminate traffic flow. We considered two classes: congested flow and free flow. Using a 1-min input signal, the DNN estimated the vehicle speed, and the 5-min average speed was calculated based on the DNN's minute-by-minute estimation. If the average speed was less than a threshold of $40\,\mathrm{km/h}$, we classified it as congested flow; otherwise, we classified it as free flow. The average speeds were obtained by calculating the harmonic mean of the DNN estimations. In this experiment, we followed the practices used on intercity highways in Japan for traffic flow discrimination thresholds and time intervals for the average speed calculation. For the evaluation, we used the two-day data obtained on October 12 and 21 as in Section V-B.

2) Experimental Results: Fig. 5 shows the confusion matrix of traffic flow discrimination between congested and free flows. Also, Table III shows scores of typical objective metrics, i.e., the accuracy, precision, recall, and F1-score. In particular, precision represents the proportion of actually congested flow data among the estimated congested flow, while recall represents the proportion of estimated congested flow data among the actual congested flow. From Fig. 5 and Table III, it can be seen that the estimator trained by the proposed method can discriminate congested flows from free flows with high accuracy.

VI. Conclusion

In this paper, a CNN was trained to estimate the total number of vehicles passing, the number of large vehicles passing, speed, and OCC from the running-vehicle sound using the traffic condition data acquired by traffic detectors installed on expressways, and the results were evaluated. A comparison of the input features showed that the best performance was achieved when the power spectrogram and phase difference were combined. The results of comparing the proposed method with conventional methods and evaluation experiments confirmed that our proposed method can estimate the total number of vehicles passing and their speeds with high accuracy. In addition, the proposed method showed high discrimination accuracy in an experiment in which we applied our proposed method to traffic flow discrimination conducted to evaluate its performance, and its practicality was confirmed.

Future work includes data augmentation and domain adaptation to improve the robustness to differences in the environment in which running-vehicle sounds are recorded during training and inference.

TABLE III
Traffic Flow Discrimination Performance

Accuracy	Precision	Recall	F1-score
95.8%	88.0%	95.7%	91.7%

References

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

[2] P. T. Martin, Y. Feng, and X. Wang, "Detector technology evaluation," Mountain-Plains Consortium (MPC), MPC-03-154, 2003.

[3] B. Coifman and S. Neelisetty, "Improved speed estimation from single-loop detectors with high truck flow," in *Journal of Intelligent Transportation Systems*, vol. 18, no. 2, pp. 138–148, 2014.

[4] M. A. Sobreira-Seoane, A. Rodríguez Molares, and J. L. Alba Castro, "Automatic classification of traffic noise," in *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3823–3823, 2008.

[5] G. Szwoch and J. Kotus, "Acoustic detector of road vehicles based on sound intensity," in *Sensors*, vol. 21, no. 23, 2021.

[6] J. Kotus and G. Szwoch, "Estimation of average speed of road vehicles by sound intensity analysis," in *Sensors*, vol. 21, no. 16, 2021.

[7] T. Toyoda, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Traffic monitoring with ad-hoc microphone array," in *Proc. 14th Inter. Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 318–322, 2014.

[8] S. Ishida, K. Mimura, S. Liu, S. Tagashira, and A. Fukuda, "Design of simple vehicle counter using sidewalk microphones," in *Proc. ITS European Congress*, pp. 1–10, 2016.

[9] A. Y. Nooralahiyan, M. Dougherty, D. McKeown, and H. R. Kirby, "A field trial of acoustic signature analysis for vehicle classification," in *Transportation Research Part C: Emerging Technologies*, vol. 5, no. 3, pp. 165–177, 1997.

[10] J. George, L. Mary, and Riyas. K. S, "Vehicle detection and classification from acoustic signal using ANN and KNN," in *Proc. Inter. Conf. on Control Communication and Computing (ICCC)*, pp. 436–439, 2013.

[11] V. Tyagi, S. Kalyanaraman, and R. Krishnapuram, "Vehicular traffic density state estimation based on cumulative road acoustics," in *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1156–1166, 2012.

[12] T. Shinohara, Y. Wakabayashi, R. Scheibler, N. Ono, N. Aizawa, and H. Nakagawa, "Sound-based speed estimation using a neural network," in *Proc. Spring Meeting of the Acoustical Society of Japan*, pp. 385–386, 2020. (in Japanese)

[13] A. Trockman and J. Z. Kolter, "Patches are all you need?" arXiv preprint arXiv:2201.09792, 2022.

[14] R. Baidya and H. Jeong, "Yolov5 with convmixer prediction heads for precise object detection in drone imagery," in *Sensors*, vol. 22, no. 21, p. 8424, 2022.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Inter. Conf. on Learning Representations*, 2015.

[16] G. Jacovitti and G. Scarano, "Discrete time techniques for time delay estimation," in *IEEE Trans. on Signal Process.*, vol. 41, no. 2, pp. 525–533, 1993.