Deep Acoustic Vehicle Counting Model with Short-Time Homomorphic Deconvolution

Technical Report

Yeonseok Park, TaeWoon Yeo, Baeksan On

KT Corporation, South Korea
{yeonseok.park, taewoon.yeo, baeksan.on}@kt.com

ABSTRACT

In the design of urban traffic monitoring solutions aimed at optimizing logistics infrastructure, acoustic vehicle counting models have gained attention for their cost-effectiveness and energy efficiency. While deep learning has proven effective in visual traffic monitoring, its application in the auditory domain remains underexplored due to the limited availability of real-world data. This study proposes the use of Short-Time Homomorphic Deconvolution (STHD) for analyzing sound signals to estimate the direction of vehicle sounds. This algorithm calculates distances between microphones based on sound direction, facilitating the inference of sound direction and movement. We present a strategy for designing and training a deep learning model that leverages features derived from this algorithm. The proposed system simultaneously counts cars and commercial vehicles on a two-lane road under moderate traffic density conditions, accurately identifying their directions of travel.

Index Terms— DCASE2024, Acoustic-Based Traffic Monitoring, Deep Acoustic Vehicle Counting Model, Short-Time Homomorphic Deconvolution

1. INTRODUCTION

The research on acoustic vehicle counting (AVC) using AI is a rapidly developing field with significant potential [1]. Although the existing literature is relatively limited, various approaches for acoustic vehicle detection [2-6] and counting [7, 8, 9, 10] have been proposed. These methods primarily rely on the analysis of audio signals captured by either a single microphone or a microphone array and utilize traditional signal processing, deep learning, or a combination of both. The basic objective is to detect passing vehicles, but more advanced techniques aim to differentiate vehicle types (e.g., cars, trucks, motorcycles), determine the direction of movement (e.g., left-to-right or right-to-left), and estimate speed.

Despite the demonstrated effectiveness of data-driven approaches in acoustic detection tasks, their potential in the field of AVC has not been thoroughly explored. One significant challenge is the scarcity of available datasets for this purpose. Existing datasets are often small, insufficient for training end-to-end deep learning models, and typically consist of single-channel recordings. The process of data collection itself is complex and costly, involving not only audio recordings but also the collection of ground truth data using other sensor modalities. Additionally, developing synchronization strategies to align the collected audio and ground truth data adds to the complexity and expense.

In this paper, we address these challenges by proposing a novel approach for AVC that leverages a multi-channel microphone array and advanced deep learning techniques. Our method includes the use of synthetic data to supplement limited real-world data, thereby improving model performance and reducing the reliance on extensive data collection efforts. This approach aims to advance the field of AVC by providing a more robust and scalable solution for accurate vehicle counting and classification.

Sound source localization is the study of classifying sound events and detecting their direction. In particular, the Homomorphic Deconvolution (HD) algorithm [11-13] is effective in estimating the source location by considering the time of flight between microphones based on the direction of the sound. As vehicles move, the position of the sound source changes relative to the microphone array. This results in variations in HD values over time, which can be used to determine the direction of movement.

In this paper, we address AVC using a 4-channel linear microphone array deployed on the side of a two-lane road. Our goal is to identify vehicle pass-by events using only 1-minute segments of sound data, leveraging efficient signal feature computation. For each detected event, the vehicle must be classified as either a car or a commercial vehicle (CV, including large vehicles such as trucks and buses) and the direction of transit must be identified. In this study, we introduce a convolutional recurrent neural network (CRNN) for traffic counting, trained using a procedure based on synthetic data generated from simulation sound data. This method significantly reduces the amount of real-world data required to achieve high traffic counting accuracy. In summary, we first define a strategy to synthesize acoustic traffic noise, then pre-train the model based on the synthetically generated dataset, and finally perform fine-tuning using a limited amount of real-world data. We demonstrate that the proposed method can effectively count traffic using only 1-minute sound data segments.

2. FEATURE EXTRACTION

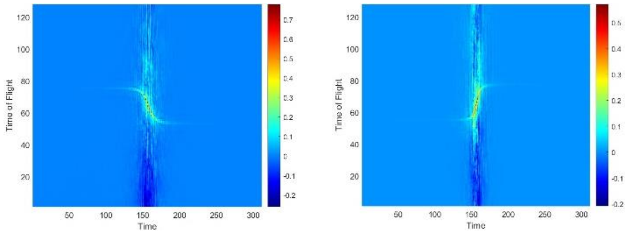
2.1 Spectrogram

A standard signal processing preprocessing algorithm, spectrogram computation, was applied to each channel of the signal. Since vehicle sounds exhibit changes at relatively low frequencies, the y-axis was cropped to the range of 1 to 128, corresponding to frequencies from 0 to 2000 Hz. This resulted in feature data of size 128x1874x4.

The spectrogram allows us to observe representative patterns in the simulation sound. When a vehicle approaches, the frequency increases, and when it moves away, the frequency decreases. This phenomenon is due to the Doppler effect. Additionally, the difference between cars and commercial vehicles can be seen in the pattern of decreasing frequencies. Cars tend to show more frequent decreases in frequency compared to commercial vehicles.

2.2 Short-Time Homomorphic Deconvolution (STHD)

in this section, we propose the Short-Time Homomorphic Deconvolution (STHD) algorithm and apply a method for computing 2D data, introducing a solution that uses these features to determine the direction of vehicle movement solely from acoustic signals. The input consists of pairs of signals summed together, resulting in 6-channel input data. When summing the signal pairs, an arbitrary delay is added to one of the channels to center the STHD algorithm's pattern. To extract the central pattern, the y-axis is cropped to the range of 129 to 384, creating feature data of size 311x128x6.



(a)

(b)

Figure 1. Results of Short-Time Homomorphic Deconvolution on Simulation Data. (a) Car moving from left to right. (b) Car moving from right to left.

Figure 1 shows the STHD results of simulation data. Figure 1 (a) illustrates the sound data when a car moves from left to right relative to the microphone array, displaying a pattern descending from top to bottom along the y-axis in the STHD output. Figure 1 (b) illustrates the sound data when a car moves from right to left relative to the microphone array, showing a pattern ascending from bottom to top along the y-axis in the STHD output.

2.3 Feature Analysis

Using feature preprocessing methods, we describe the actual data analysis of "loc1/train/00000.flac". The labels for this sound data are as shown in Table 1.

Table 1. label of loc1/train/00000.flac

Label Name	car-l2r	car-r2l	CV-l2r	CV-r2l
Label (counts)	1	1	1	0

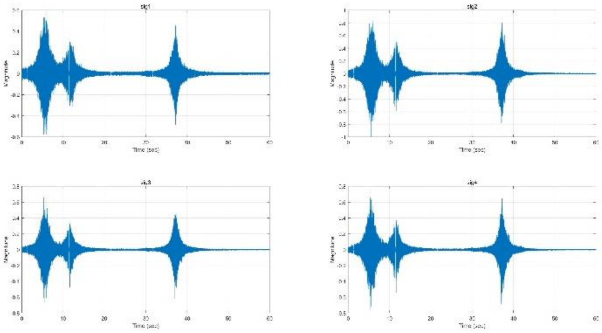


Figure 2. Raw signal 4-channel graph of "loc1/train/00000.flac".

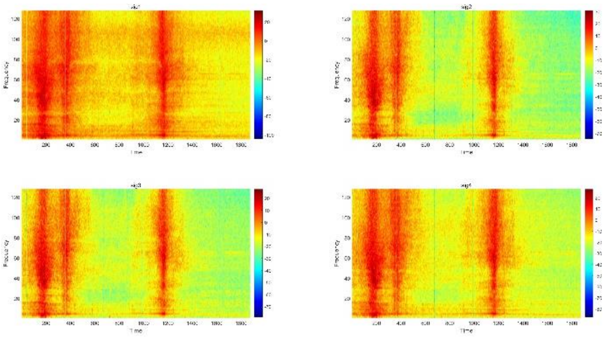


Figure 3. Spectrogram features of 4 channels for "loc1/train/00000.flac".

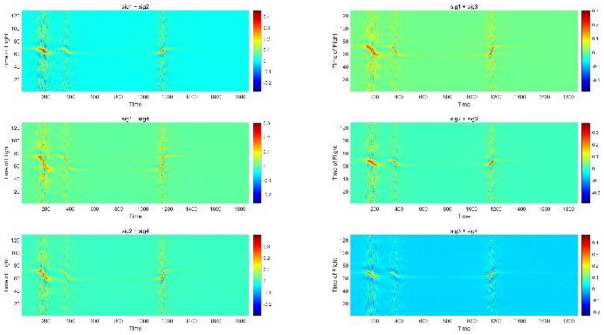


Figure 4. STHD features of 6 channels for "loc1/train/00000.flac".

Especially in Figure 4, multiple vehicles moving can be observed, showing patterns both upwards and downwards along the time axis. There are two instances of downward patterns, indicating vehicles moving from left to right: one labeled as car-l2r and one as CV-l2r. Additionally, one instance of an upward pattern is visible, indicating a vehicle moving from right to left, labeled as car-r2l. This feature extraction method proposed enables precise analysis of vehicle types and their directional movements.

3. MODEL ARCHITECTURE

2.2. CRNN Architecture

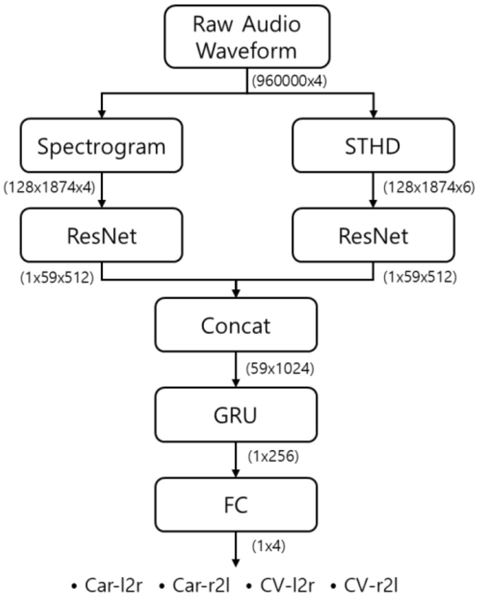


Figure 5. The proposed CRNN architecture takes the raw signal input from a 4-channel linear microphone array and computes: a) a spectrogram; b) the STHD with the sum of pairs of channels. Two ResNet models compute sound information and sound movement information features, respectively. The concatenated features are processed along the time dimension, followed by a Gated Recurrent Unit (GRU) and a fully connected (FC) layer for regressing the number of vehicles per type (car, CV) and per direction (left-to-right, right-to-left).

The the convolutional recurrent neural network (CRNN) architecture used in this study integrates CNNs to learn patterns in features related to vehicle sound information and direction. The outputs from CNNs provide separate predictions for sound characteristics and movement directions of vehicles. RNNs are employed to learn temporal dependencies in both sound characteristics and directional information, enabling the model to infer which vehicles are moving in which direction over time.

The CNN component utilizes the ResNet18 architecture. The first layer consists of Conv2D with 64 filters, tailored to accommodate the input channels of each feature type. Specifically, Spectrogram features are input with 4 channels, while STHD features are input with 6 channels. The final global average pooling layer is modified to fix the time dimension, enhancing the capture of temporal data flow by the Gated Recurrent Unit (GRU).

The outputs from each feature type are concatenated. The final fully connected (fc) layer of ResNet18 is removed, and the data is reshaped to (59x1024) along the time dimension, which conforms to the sequential data format required by the GRU.

The GRU has a hidden dimension of 256 and consists of 2 layers. Specifically, it employs a non-directional GRU to process sequential data. This choice is made because the patterns in STHD features resemble horizontally flipped patterns of different labels (e.g., the temporal pattern of car-l2r is similar to the horizontally flipped temporal pattern of car-r2l), potentially causing confusion if bidirectional GRU were used.

Finally, a FC layer with 4 neurons serves as the regression output, providing vehicle counts in four categories: car-l2r, car-r2l, cv-l2r, cv-r2l. The model comprises a total of 23.7M trainable parameters.

4. EXPERIMENTAL RESULTS

4.1 Synthetic Simulation Data

One of the primary challenges in deep learning models is the requirement for large volumes of training data. To address this, we synthesized a new dataset of 6000 samples by augmenting publicly available simulation data. Labels were randomly generated, with a maximum of 20 cars and 5 commercial vehicles (CVs) per sample. The synthesized sound data was generated using MATLAB.

4.2 Experimental Setup

In this section, we detail the experimental setup conducted for training and evaluating our model.

The sample frequency is throughout 16 kHz in all data and applying peak normalization to each audio segment. During training, we employed a mean squared error loss function with the Adam optimizer set to a learning rate of lr = 10−4 and a batch size of 16. The model was trained for 100 epochs, with the best checkpoint selected based on validation loss criteria.

The proposed architecture was implemented using PyTorch and trained on a single RTX 4060Ti-16GB GPU to leverage computational efficiency. Initially, we trained the model using only Synthetic Simulation Data, dividing it into separate train and validation sets for pre-training the model. Subsequently, we applied transfer learning using the pre-trained model on real-world data. Transfer learning was conducted individually for each location (loc1 ~ 6) to adapt the model to specific environmental variables surrounding each recording location, such as ambient noise conditions and microphone array characteristics.

5. CONCLUSION

In this study, we propose an AVC system based on a CRNN architecture utilizing a novel feature extraction method, including Short-Time Homomorphic Deconvolution. The model aims to count traffic for various vehicle types and directions of transit. Pre-training on synthesized simulation data and transfer learning on real data provide significant advantages in achieving high counting accuracy. Future research will analyze the impact of non-vehicle sounds and microphone array noise on counting performance and evaluate the model's generalization capabilities across diverse environments. Parts of this work have been submitted for patent approval or are currently in progress.

6. REFERENCES

[1] S. Damiano, L. Bondi, S. Ghaffarzadegan, A. Guntoro, and T. van Waterschoot, “Can Synthetic Data Boost the Training of Deep Acoustic Vehicle Counting Networks?,” arXiv:2401.09308, 2024

[2] S. Ishida, J. Kajimura, M. Uchino, S. Tagashira, and A. Fukuda, “SAVeD: Acoustic vehicle detector with speed estimation capable of sequential vehicle detection,” in 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 906-912, IEEE, November 2018

[3] C. Wang, Y. Song, H. Liu, J. Liu, H. Liu, B. Li, and X. Yuan, “Real-time vehicle sound detection system based on depthwise separable convolution neural network and spectrogram augmentation,” Remote Sensing, vol. 14, no. 19, p. 4848, 2022.

[4] G. Szwoch and J. Kotus, “Acoustic detector of road vehicles based on sound intensity,” Sensors, vol. 21, no. 23, p. 7781, 2021.

[5] K. Kubo, C. Li, S. Ishida, S. Tagashira, and A. Fukuda, “Design of ultra low power vehicle detector utilizing discrete wavelet transform,” in Proc. ITS AP Forum, pp. 1052-1063, May 2018.

[6] N. Bulatović and S. Djukanović, “Mel-spectrogram features for acoustic vehicle detection and speed estimation,” in 2022 26th International Conference on Information Technology (IT), pp. 1-4, IEEE, February 2022.

[7] S. Djukanović, J. Matas, and T. Virtanen, “Robust audio-based vehicle counting in low-to-moderate traffic flow,” in 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 1608-1614, October 2020.

[8] S. Djukanović, Y. Patel, J. Matas, and T. Virtanen, “Neural network-based acoustic vehicle counting,” in 2021 29th European Signal Processing Conference (EUSIPCO), pp. 561-565, August 2021.

[9] X. Zu, S. Zhang, F. Guo, Q. Zhao, X. Zhang, X. You, et al., “Vehicle counting and moving direction identification based on small-aperture microphone array,” *Sensors*, vol. 17, no. 5, p. 1089, 2017.

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

[11] Y. Park, A. Choi, and K. Kim, “Monaural sound localization based on reflective structure and homomorphic deconvolution,” *Sensors*, vol. 17, no. 10, p. 2189, 2017.

[12] Y. Park, A. Choi, and K. Kim, “Parametric Estimations Based on Homomorphic Deconvolution for Time of Flight in Sound Source Localization System,” *Sensors*, vol. 20, no. 3, p. 925, 2020.

[13] Y. Park, A. Choi, and K. Kim, “Single-Channel Multiple-Receiver Sound Source Localization System with Homomorphic Deconvolution and Linear Regression,” *Sensors*, vol. 21, no. 3, p. 760, 2021.