JLESS SUBMISSION TO DCASE2024 TASK10: AN ACOUSTIC-BASED TRAFFIC MONITORING SOLUTION

Technical Report

Dongzhe Zhang[1,2], Jisheng Bai[1,2], Jianfeng Chen[1,2]

[1] Joint Laboratory of Environmental Sound Sensing,
School of Marine Science and Technology,
Northwestern Polytechnical University, Xi'an, China
[2] LianFeng Acoustic Technologies Co., Ltd. Xi'an, China
{dongzhezhang2022, baijs}@mail.nwpu.edu.cn, chenjf@nwpu.edu.cn

ABSTRACT

In this technical report, we describe our proposed system for the traffic monitoring challenge. Our solution addresses the critical need for efficient traffic monitoring systems in smart city development, leveraging the advantages of acoustic sensors. Initially, we review various sensor types used in traffic monitoring, emphasizing the benefits of acoustic sensors such as low cost, power efficiency, and robustness in adverse conditions. Given the challenges of collecting and labeling real-world traffic data, we incorporate synthetic data generated via the pyroadacoustics simulator to enhance system performance. We employ multiple data augmentation techniques to create a balanced and comprehensive training dataset. Our approach also includes detailed metadata integration, which provides sensor location IDs, timestamps, sensor array geometry, and vehicle counts. During the training phase, we implement several strategies to improve the system's generalization in real-world environments. Our results demonstrate that the proposed system significantly outperforms baseline models in accurately detecting and classifying traffic events, validating the efficacy of our approach using both real and synthetic data.

***Index Terms*—**

1. INTRODUCTION

Traffic monitoring solutions are essential for smart city development, helping to monitor roadway infrastructure usage and conditions and detect anomalies [1]. These systems use various sensors categorized into two main types: intrusive and non-intrusive sensors. Intrusive sensors, such as induction loops and vibration or magnetic sensors, are embedded directly into the road. Non-intrusive systems, like radar, cameras, infrared or acoustic sensors, and off-road mobile devices such as aircraft or satellites, are mounted over or on the side of the road. Acoustic sensors, in particular, offer several advantages that make them a desirable choice either alone or combined with other sensors. Their benefits include low cost, power efficiency, ease of installation, and robustness to adverse weather and low-visibility conditions. Collecting and labeling real-world traffic data is challenging and resource-intensive. To address this issue, this challenge investigates the impact of using synthetic data generated by traffic simulators on system performance. Alongside real-world traffic sound recordings, a tool is provided to synthesize realistic vehicle pass-by events under various traffic conditions using the open-source pyroadacoustics road acoustics simulator [2].

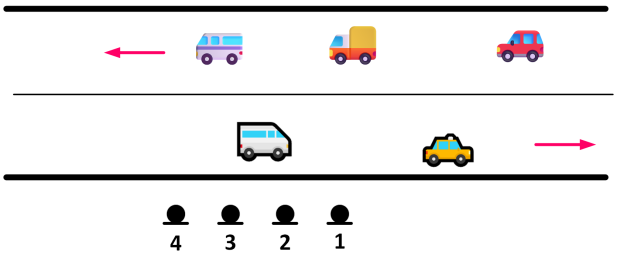

Figure 1: System overview

2. PROPOSED METHOD

In this section, we first introduce the input features of the proposed system. Then we introduce the network architecture and training procedures.

2.1. Features

The primary input for the model is a batch of raw audio waveforms. The transformation of these raw signals into meaningful features is pivotal for effective traffic monitoring. Initially, the raw waveform is converted into its time-frequency representation using the Short-Time Fourier Transform (STFT). The STFT decomposes the signal into its constituent frequencies over short time intervals, facilitating the analysis of its spectral content. To enhance the interpretability and relevance of the spectral features, the magnitude of the STFT output is mapped onto the mel scale, resulting in a log-mel spectrogram. This transformation aligns the frequency representation more closely with human auditory perception, emphasizing perceptually significant frequencies. Concurrently, the Generalized Cross-Correlation (GCC) is computed from the STFT output. The GCC captures the time delay between audio channels, providing critical spatial information that aids in direction-of-arrival (DOA) estimation and sound source localization. Both log-mel and GCC features undergo normalization to ensure uniform feature scales, which is

essential for stable and efficient model training. The final step involves concatenating the log-mel and GCC features, creating a comprehensive feature set that leverages both spectral and spatial information. This combined feature representation enhances the model's ability to accurately classify different audio events, leading to robust traffic monitoring and anomaly detection.

2.2. Network architecture

Following the temporal merging, the data is processed through a series of transformer encoder layers. Transformer encoders are particularly effective for sequence modeling tasks due to their self-attention mechanism, which allows them to capture long-range dependencies and contextual information more efficiently than recurrent neural networks (RNNs). The self-attention mechanism in transformer encoders can simultaneously attend to information from different positions in the sequence, making it well-suited for understanding complex temporal relationships within the audio signal. By replacing the RNN layer with transformer encoder layers, the model gains several advantages. Transformer encoders can process sequences in parallel rather than sequentially, leading to improved computational efficiency and scalability. Additionally, the self-attention mechanism enables the model to weigh the importance of different time steps more effectively, enhancing its ability to capture intricate temporal dependencies that are crucial for accurately distinguishing between different types of vehicles and their directions of travel. This is particularly beneficial for traffic monitoring, where the audio signal may contain overlapping sounds from multiple vehicles and environmental noise. The output from the transformer encoder layers provides a refined feature set that encapsulates both the temporal dynamics and the spatial information derived from the earlier stages. This refined feature representation is then utilized in the subsequent regression layer to predict specific vehicle counts. The final stage of the model involves a regression layer that takes the output from the last transformer encoder layer. This step is critical for translating the processed features into specific vehicle counts. The regression layer outputs a set of values corresponding to the counts of four different classes: cars traveling to the right, cars traveling to the left, commercial vehicles (CV) traveling to the right, and commercial vehicles traveling to the left. This output format directly addresses the traffic monitoring objectives by providing precise counts for each vehicle class and direction.

3. EXPERIMENTS

3.1. Experimental settings

We adhere to the baseline settings for feature extraction, employing a sampling frequency of 16kHz, 128 Mel filters, and an STFT configuration with a frame length of 64ms and a frame hop of 10ms. Our training process begins with a batch size of 64 and involves two stages: initially training on simulated data for 100 epochs with a learning rate of 0.005. In the fine-tuning stage, we further refine the model using real recordings for an additional 100 epochs. During this stage, the learning rate is adjusted to 0.1, with decay applied to optimize performance.

3.2. Results

There are two evaluation metrics: The RMSE value is influenced by the range of data values: since the variation range of CV numbers is not large, RMSE tends to be lower overall. Kendall's Tau Corr, primarily emphasizes the sequential correlation between two sequences and is less affected by numerical values. This aspect is crucial for monitoring traffic flow trends. Table 1 shows the performance of the development set for the proposed methods.

Table 1: Results on Different Locations

Location	Metric	car_left	car_right	cv_left	cv_right
loc1	Kendall's Tau Corr	0.437	0.44	0.149	0.128
	RMSE	2.571	2.928	0.905	0.889
loc2	Kendall's Tau Corr	0.53	0.461	0.109	0.134
	RMSE	3.196	2.446	0.89	0.658
loc3	Kendall's Tau Corr	0.543	0.566	0.191	0.32
	RMSE	1.744	1.309	0.314	0.213
loc4	Kendall's Tau Corr	0.463	-0.063	-0.254	-0.261
	RMSE	1.653	1.228	0.934	0.515
loc5	Kendall's Tau Corr	0.546	0.445	-0.08	-0.045
	RMSE	0.708	0.717	0.385	0.279
loc6	Kendall's Tau Corr	0.806	0.714	0.736	0.688
	RMSE	1.598	1.742	0.544	0.52

4. CONCLUSION

In this report, we present our system designed for efficient traffic monitoring in smart city development, leveraging acoustic sensors. We review sensor types, emphasizing acoustic sensors' cost-effectiveness, energy efficiency, and resilience. To address data challenges, we augment real-world data with synthetic data from the pyroadacoustics simulator. Our exploration of various network architectures resulted in performance improvements over baseline methods.

5. REFERENCES

[1] S. Damiano and T. van Waterschoot, "Pyroadacoustics: a Road Acoustics Simulator Based on Variable Length Delay Lines," in *Proceedings of the 25th International Conference on Digital Audio Effects (DAFx20in22)*, Vienna, Austria, September 2022, pp. 216–223.

[2] S. Damiano, L. Bondi, S. Ghaffarzadegan, A. Guntoro, and T. van Waterschoot, "Can synthetic data boost the training of deep acoustic vehicle counting networks?" in *Proceedings of the 2024 International Conference on Acoustics, Speech and Signal Processing (ICASSP) (accepted)*, Seoul, South Korea, April 2024.