FINE-GRAINED AUDIO FEATURE REPRESENTATION WITH PRETRAINED MODEL AND GRAPH ATTENTION FOR TRAFFIC FLOW MONITORING

Technical Report

Shitong Fan[1]*, Feiyang Xiao*[1]*, Shuhan Qi*[2]*, Qiaoxi Zhu*[3]*, Wenwu Wang*[4]*, and Jian Guan*[1*]

[1]Group of Intelligent Signal Processing (GISP), College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
[3]University of Technology Sydney, Ultimo, Australia
[4]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK

ABSTRACT

This technical report describes our submission for DCASE 2024 Challenge Task 10. To enhance audio feature representation for audio event detection, we use pre-trained audio neural networks (PANNs) for audio feature pretraining and a graph attention module (GAT) for audio feature fine-tuning to capture important temporal relations and learn the dependencies among audio features across different time frames. Thus, our method can capture important audio event information in the audio signals, and provide fine-grained audio representation for vehicle type detection. We use this fine-grained feature instead of the feature branch in the original baseline to build our systems. In our systems, we apply the SpecAugment strategy for audio data augmentation and introduce an overall phase shift to explore the directional information. Experimental results indicate that our systems show some improved performance among the six locations in the evaluation, except for location 1.

***Index Terms*—** Acoustic-based traffic monitoring, transfer learning, pretrained model, graph attention network, phase shift

1. INTRODUCTION

The development of smart cities relies on deploying sensors and devices in urban areas to collect data, enabling effective monitoring and management of public infrastructure [1]. Among various types of sensors, acoustic sensors offer advantages compared to others due to their low cost, power efficiency, ease of installation, and resilience to adverse weather conditions and low visibility. This makes it an ideal choice for use independently or in combination with other sensors. Therefore, monitoring traffic flow through sounds collected from road scenes has significant practical relevance. Task 10 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2024 Challenge introduces an acoustic-based traffic monitoring system [2, 3]. This task aims to design a solution that utilizes acoustic signals to count the number of vehicles, distinguish between vehicle types (such as cars or commercial vehicles), and record their travel direction (left turn or right turn).

This technical report presents our submission systems for Task 10. Our systems enhance the detection of vehicle types passing by in audio signals by exploiting the temporal relation with audio pretraining and graph attention, to capture audio event information, thus improving overall performance. During the implementation of our systems, we leverage our previous work, GraphAC [4], to develop our system and introduce PANNs [5] and GAT [6] to model the temporal relation of audio signals.

Additionally, we apply SpecAugment [7] to augment audio data, and exploit the directional information of passing vehicles by processing the phase information from four channels.

2. OUR SYSTEMS

We improve the baseline method by replacing its feature extraction branch with fine-grained feature representation using PANNs and GAT, forming our submission System-1. Specifically, we use pretrained PANNs to extract audio features from the audio signals. The audio feature frames extracted by PANNs are then treated as graph nodes, and GAT is adopted to exploit the relation between feature nodes to learn the dependencies among audio feature frames, achieving more fine-grained audio representations for audio event detection (i.e., vehicle types detection).

After obtaining the fine-grained audio features for vehicle types detection, we use the same way as baseline method to extract features related to the vehicle's movement direction. We then fuse the fine-grained feature and the feature for vehicle movement direction to obtain a comprehensive representation for both vehicle's type and its direction of travel. Here, due to the inconsistency temporal dimension between fine-grained feature and the vehicle movement direction feature, we apply temporal average pooling to the vehicle movement direction feature. This facilitates the fusion of features along the corresponding temporal dimension, resulting in the feature representation that contains both vehicle type information and movement direction information at each time step, and obtain our submission System-1.

Based on System-1, we apply SpecAugment [7] for data augmentation to build the System-2, and introduce an overall phase shift to build the System-3. System-2 uses the SpecAugment strategy to augment the audio data for detection while maintaining the same model structure as System-1, whereas System-3 introduces an overall phase shift to account for the direction of the passing vehicle. Specifically, System-3 uses the short-time Fourier transform (STFT) to extract phase information. By taking the phase of the first channel as the reference phase, we have the phase shift of other

*Corresponding author.

Table 1: Performance comparison in terms of Kendall's Tau Rank Correlation (Kendall's Tau Corr) and Root Mean Square Error (RMSE) across different sound events and locations in the development dataset of DCASE 2024 challenge Task 10.

(a) Performance on the Sound Event of **"car_left"**

Methods	loc1		loc2		loc3		loc4		loc5		loc6	
	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE
Baseline	**0.470**	**2.449**	0.446	3.308	**0.619**	**1.629**	0.456	1.698	0.484	**0.662**	**0.825**	1.672
System-1	0.424	2.575	0.706	2.219	0.539	1.761	**0.512**	**1.548**	0.530	0.741	0.808	1.662
System-2	0.434	2.600	**0.719**	**2.177**	0.551	1.729	0.097	2.095	**0.557**	0.708	0.816	**1.582**
System-3	0.437	2.567	0.478	3.335	0.531	1.780	0.122	2.009	0.226	0.875	0.520	3.430

(b) Performance on the Sound Event of **"car_right"**

Methods	loc1		loc2		loc3		loc4		loc5		loc6	
	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE
Baseline	**0.478**	**2.687**	0.221	3.560	**0.593**	**1.209**	0.248	2.210	**0.575**	**0.607**	**0.736**	1.950
System-1	0.447	2.955	**0.497**	**2.347**	0.573	1.285	**0.391**	**1.217**	0.363	0.738	0.672	1.983
System-2	0.448	2.919	0.401	2.666	0.577	1.275	0.240	1.548	0.401	0.697	0.684	**1.910**
System-3	0.459	2.874	0.266	3.162	0.554	1.319	-0.240	1.572	-0.159	0.844	0.520	2.722

(c) Performance on the Sound Event of **"cv_left"**

Methods	loc1		loc2		loc3		loc4		loc5		loc6	
	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE
Baseline	**0.231**	**0.732**	0.135	**0.468**	0.102	0.308	0.000	**0.548**	0.092	0.491	0.711	**0.535**
System-1	0.173	0.868	0.116	0.818	0.146	**0.306**	-0.383	0.741	0.022	0.398	**0.725**	0.583
System-2	0.207	0.892	0.226	0.783	**0.171**	0.315	**0.182**	0.604	0.058	0.362	0.683	0.604
System-3	0.176	1.007	**0.252**	0.815	0.084	0.311	-0.383	0.679	**0.108**	**0.357**	0.406	0.940

(d) Performance on the Sound Event of **"cv_right"**

Methods	loc1		loc2		loc3		loc4		loc5		loc6	
	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE	Kendall's Tau Corr	RMSE
Baseline	**0.189**	**0.777**	-0.026	0.610	0.272	0.199	0.438	0.728	0.108	0.676	**0.648**	**0.441**
System-1	0.144	0.891	**0.283**	**0.594**	0.356	0.206	**0.562**	0.398	0.193	0.287	0.584	0.599
System-2	0.126	0.861	0.171	0.677	**0.377**	**0.195**	0.445	0.428	**0.357**	**0.208**	0.570	0.594
System-3	0.159	0.940	-0.128	0.731	0.277	0.227	0.531	**0.387**	0.283	0.272	0.252	0.795

channels, and the overall phase shift are calculated as a feature to determine the direction of vehicle movement in our System-3.

3. EXPERIMENTS

3.1. Dataset

We conduct experiments on the development dataset of the DCASE 2024 Challenge Task 10. This dataset includes real data recordings from six different locations (loc1 to loc6) as well as simulated data. The real data recordings are captured by a linear microphone array installed parallel to the direction of traffic flow on the side of the road. Due to the difficulties associated with collecting and labeling real traffic data, the dataset also includes simulated data generated by an acoustic traffic simulator [3], which simulates the sounds of vehicles moving along arbitrary trajectories and speeds in real-world scenarios. The dataset describes sound events as "car_left", "car_right", "cv_left" and "cv_right". For example, "car_left" indicates an event where a car is moving from left to right. Similar to the baseline approach, we include synthetic data in the training process. We train and validate our proposed systems on the development dataset to verify the effectiveness of our method.

3.2. Experimental Setup

For the PANNs module [5], we select the CNN10 architecture and load the corresponding pretrained weights. The proposed methods are evaluated on a single NVIDIA 3090 GPU, using the same learning rate and learning rate update strategy as the baseline configuration [3].

3.3. Evaluation Metric

According to the metrics specified by the competition website[1], we use Kendall's Tau Rank Correlation (Kendall's Tau Corr) and Root Mean Square Error (RMSE) for evaluation. Here, Kendall's Tau Corr represents the correlation between the predicted results and the actual results, while RMSE represents the prediction error.

3.4. Results

We compare our systems with the baseline system of Task 10. The results are given in Table 1, where we can see that System-1 and System-2 achieve performance improvement over the baseline system, i.e., at loc2, loc3, loc4, loc5 and loc6, which verify the effectiveness of the SpecAugment. Although System-3 performs worse than other systems, it provides improvement for certain locations and categories, which shows that overall phase shift may be useful for vehicle movement direction.

4. CONCLUSION

We introduce our submission systems by introducing a fine-grained feature representation method using audio pretraining and GAT for DCASE 2024 Challenge Task 10, and the experimental results show that our submission systems outperform the baseline system to some extent across multiple locations, with a particularly notable improvement at loc2. Furthermore, we also explore the possibility of utilizing overall phase shift information to detect the driving direction of vehicles in System-3.

[1] https://dcase.community/challenge2024/task-acoustic-based-traffic-monitoring

5. REFERENCES

[1] R. Du, P. Santi, M. Xiao, A. V. Vasilakos, and C. Fischione, "The Sensable City: A survey on the deployment and management for smart city monitoring," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1533–1560, 2018.

[2] S. Damiano and T. van Waterschoot, "Pyroadacoustics: A road acoustics simulator based on variable length delay lines," in *Proc. of International Conference on Digital Audio Effects (DAFx20in22)*, September 2022, pp. 216–223.

[3] 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 *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2024.

[4] F. Xiao, J. Guan, Q. Zhu, and W. Wang, "Graph attention for automated audio captioning," *IEEE Signal Processing Letters*, vol. 30, pp. 413–417, 2023.

[5] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[6] P. Casanova, A. R. P. Lio, and Y. Bengio, "Graph attention networks," *ICLR. Petar Velickovic Guillem Cucurull Arantxa Casanova Adriana Romero Pietro Liò and Yoshua Bengio*, 2018.

[7] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.