Evaluation of Pitch-Shift Augmentation and Average Pooling for Acoustic Scene Classification under Unseen-City Conditions

Masayuki Sera*, Takao Kawamura*, and Nobutaka Ono*

***Abstract*—In this report, we present our submission for the APSIPA ASC 2025 Grand Challenge on semi-supervised acoustic scene classification under domain shift, focusing on the effectiveness of waveform-level pitch-shift augmentation and temporal average pooling for improving robustness to recordings from unseen cities. We employ a city-disjoint cross-validation scheme based on the official meta-data of city information, splitting the labeled development set into two folds with non-overlapping training and testing cities. We apply pitch-shift augmentation to increase data diversity and replace the baseline's temporal max pooling with average pooling to better integrate information over time. Experimental results demonstrate that both techniques improve classification accuracy for unseen cities, and that their combination achieves a macro-average accuracy of 44.17%, representing a 4.8-point gain over the baseline.**

I. Introduction

Acoustic scene classification (ASC) aims to categorize an audio recording into a predefined set of scene classes (e.g., "airport", "restaurant", "park") based on the overall acoustic characteristics of the recording [1], [2]. It has applications in environmental monitoring, smart cities, and context-aware services [3]. While deep learning methods have greatly advanced ASC performance, a persistent challenge is domain shift, where performance drops when test data differ from training data, such as recordings from unseen cities or devices [4]–[6].

The APSIPA ASC 2025 Grand Challenge [7] provides a suitable testbed for addressing this issue, offering labeled and unlabeled audio from multiple cities and enabling explicit evaluation under unseen-city conditions. Beyond architectural innovations, data augmentation is a practical and widely adopted strategy to improve robustness to unseen data in ASC. For example, Salamon and Bello [8] demonstrated that spectral transformations such as pitch shifting and time stretching can improve environmental sound classification performance. More recently, Li et al. [9] applied lightweight models with augmentation strategies to achieve competitive ASC performance under complexity constraints.

Among various augmentation techniques, pitch shifting is notable for its ability to alter the spectral characteristics of an audio signal while preserving its temporal structure. In this work, we examine waveform-level pitch-shift augmentation applied during training, combined with a pooling strategy modification from max pooling to average pooling in the baseline SE-Trans architecture [10]. Our goal is to evaluate whether these simple, lightweight modifications can yield measurable gains in macro-average accuracy for unseen cities without increasing model complexity.

II. Dataset Split and Baseline System

Following the challenge setting, only labeled data from the development set were used for training and evaluation. Cities were split into two folds:

- Fold 1: Train on Xi'an, Chongqing, Shangrao, Jinan; Test on Luoyang, Hefei, Shanghai, Liupanshui.
- Fold 2: Swap train and test city sets.

The distribution of labeled samples per scene and city is shown in Fig. 1, which allows explicit evaluation of generalization under unseen-city conditions. The baseline system is the SE-Trans architecture [10], which processes $T \times F$ log-mel spectrograms through two squeeze-and-excitation (SE) blocks, a Transformer encoder, temporal max pooling, and a fully connected classifier.

III. Proposed Method

A. Pitch-Shift Augmentation

To augment the training set and simulate variations in recording conditions and devices, we perform pitch shifting on a per-sample basis during training. Specifically, each waveform is pitch-shifted with a probability of 0.5, using a random shift between -2.0 and $+2.0$ semitones before feature extraction. This operation is applied only during training; validation and test samples remain unchanged. The transformation is implemented within the dataset loader using `librosa.effects.pitch_shift` [11], and is applied immediately after loading the waveform in the data loader, before computing the log-mel spectrogram features used as model input.

B. Average Pooling

We replace the temporal max pooling in the baseline with average pooling, allowing all time frames to contribute to the final representation. This aggregation can reduce sensitivity to isolated peaks and may better capture the overall temporal structure, and in our experiments it consistently improved performance over the baseline across folds.

	location	Airport	Bar	Bus	Site	Metro	Square	Restaurant	Mall	Street	Park	Total
split 1	Jinan	0	0	0	94	0	0	0	0	0	0	94
	Shangrao	0	0	100	0	0	0	0	0	0	0	100
	Chongqing	0	80	0	0	0	0	0	0	0	52	132
	Xi'an	113	0	0	0	109	174	101	81	143	55	776
split 2	Hefei	107	0	88	0	0	0	0	0	0	0	195
	Liupanshui	0	0	0	0	0	0	72	0	0	0	72
	Luoyang	0	85	0	79	0	0	0	32	0	41	237
	Shanghai	0	0	0	0	100	0	0	34	0	0	134
	Total	220	165	188	173	209	174	173	147	143	148	1740

Fig. 1. Class-wise Sample Counts for the Two-Fold Cross-Validation Setting (Labeled Data)

TABLE I
MACRO-AVERAGE ACCURACY (%) FOR UNSEEN-CITY EVALUATION ACROSS FOLDS.

Method	Fold 1	Fold 2	Average
Baseline	41.07	37.71	39.39
Pitch-shift	48.28	37.07	42.68
Average Pooling (AP)	47.81	37.96	42.89
AP + Pitch-shift	**49.22**	**39.11**	**44.17**

IV. Experimental Setup

Recordings were resampled to 44.1 kHz and converted to 64-bin log-mel spectrograms using a 40 ms Hann window and a 20 ms hop size. We trained models using the Adam optimizer (learning rate 10^{-4}, batch size 64), with early stopping after 5 epochs without validation improvement, and a maximum of 100 epochs.

V. Results

Table I shows macro-average accuracies for each method. Pitch shifting alone and average pooling alone each improved performance over the baseline, and their combination achieved the best result, 4.8 points higher than the baseline.

VI. Conclusion

We investigated waveform-level pitch-shift data augmentation and average pooling for ASC under unseen-city conditions. The combination improved macro-average accuracy from 39.39% to 44.17%, demonstrating that simple spectral augmentation and improved temporal aggregation can enhance domain generalization in ASC.

Acknowledgment

This work was supported by JST SICORP Grant Number JPMJSC2306.

References

[1] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2018, pp. 9–13.

[2] T. Heittola, A. Mesaros, and T. Virtanen, *Acoustic scene classification in DCASE 2020 challenge: Generalization across devices and low complexity solutions*, 2020. arXiv: 2005.14623 [eess.AS].

[3] B. Ding, T. Zhang, C. Wang, *et al.*, "Acoustic scene classification: A comprehensive survey," *Expert Systems with Applications*, vol. 238, p. 121 902, 2024. DOI: 10.1016/j.eswa.2023.121902.

[4] K. Drossos, P. Magron, and T. Virtanen, "Unsupervised adversarial domain adaptation based on the Wasserstein distance for acoustic scene classification," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 259–263. DOI: 10.1109/WASPAA.2019.8937231.

[5] W. Wei, H. Zhu, E. Benetos, and Y. Wang, "A-CRNN: A domain adaptation model for sound event detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 276–280. DOI: 10.1109/ICASSP40776.2020.9054248.

[6] Y. Tan, H. Ai, S. Li, and M. D. Plumbley, "Acoustic scene classification across cities and devices via feature disentanglement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1286–1297, 2024. DOI: 10.1109/TASLP.2024.3353578.

[7] J. Bai, M. Wang, H. Liu, *et al.*, *Description on IEEE ICME 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift*, 2024. arXiv: 2402.02694 [eess.AS].

[8] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[9] Y. Li, W. Cao, W. Xie, Q. Huang, W. Pang, and Q. He, "Low-complexity acoustic scene classification using data augmentation and lightweight resnet," *arXiv preprint arXiv:2306.02054*, 2023.

[10] J. Bai, J. Chen, M. Wang, M. S. Ayub, and Q. Yan, "A squeeze-and-excitation and transformer-based cross-task model for environmental sound recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 3, pp. 1501–1513, 2023. DOI: 10.1109/TCDS.2022.3222350.

[11] B. McFee, C. Raffel, D. P. W. Liang, *et al.*, "Librosa: Audio and music signal analysis in python," in *Proc. the 14th Python in Science Conference*, 2015.