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

Takao Kawamura*, Masayuki Sera*, and Nobutaka Ono*
* Tokyo Metropolitan University, Japan
E-mail: kawamura-takao@ed.tmu.ac.jp, sera-masayuki@ed.tmu.ac.jp, onono@tmu.ac.jp

***Abstract*—In this report, we describe our submitted systems for the APSIPA ASC 2025 Grand Challenge, targeting improved acoustic scene classification (ASC) with enhanced generalization across cities. To evaluate robustness to unseen cities, we adopt a city-disjoint cross-validation scheme by splitting the development dataset into two folds with non-overlapping cities. To reduce the risk of overfitting with limited labeled data, we apply two feature-level methods: restricting the input to low-frequency bands to reduce the input complexity and applying pitch-shift augmentation to increase variability. We then replace the temporal max pooling in the baseline with average pooling, allowing information from all time frames to contribute to the final representation. Experimental results show that average pooling improves classification accuracy for unseen cities, and that combining it with feature-level methods achieves about an 8-point improvement in macro-average accuracy over the baseline in the best configuration. In the APSIPA ASC 2025 Grand Challenge, the average pooling system with low-pass filtering ranked second, and the one with pitch-shift augmentation ranked fourth, demonstrating the effectiveness of our feature-level methods.**

I. Introduction

Acoustic Scene Classification (ASC) is the task of classifying an audio recording into a predefined scene label (e.g., "Airport," "Restaurant," "Park"). ASC has potential applications in life-logging, environmental monitoring, and smart home technologies [1]. It has been extensively investigated as one of the core tasks in the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge [2], [3]. It is also featured as a topic in the APSIPA ASC 2025 Grand Challenge [4].

Recent advances in ASC have been driven by deep learning methods, which have greatly improved performance. However, deep learning-based ASC models face two major challenges: domain shift [5]–[8], where mismatches between training and testing data degrade performance, and scarcity of labeled data, which limits the effectiveness of supervised training. Domain shift arises from differences in recording environments including cultural and infrastructural variations across cities, which make generalization difficult. Semi-supervised learning [9] has been explored to alleviate the lack of labeled data, but its effectiveness depends on the model performance, and evaluating generalization under domain shift remains challenging.

To evaluate robustness to domain shift, we adopt a city-disjoint cross-validation scheme, splitting the labeled development dataset into two folds with non-overlapping training and testing cities based on the provided city metadata. This scheme provides a foundation for selecting and validating robust approaches.

The characteristics of our approach are summarized as follows. First, we design a dataset split based on city metadata to evaluate model generalization to unseen cities. Second, we introduce two feature-level methods: (i) reducing the complexity of the input by using only the low-frequency range, and (ii) increasing variability through pitch-shift augmentation. Third, we replace max pooling in the baseline with average pooling to explicitly aggregate information from all frames. In evaluation experiments using the development dataset of the APSIPA ASC 2025 Grand Challenge [4], we confirm that our approach achieves an 8-point improvement over the baseline in unseen city environments.

II. Baseline Model Architecture

The architecture of the ASC model used in the baseline is the Squeeze-and-Excitation and Transformer (SE-Trans) [10]. The network architecture of the baseline is shown in Table I. In the baseline of the APSIPA ASC 2025 Grand Challenge, the baseline adopts the best configuration of the SE-Trans from [10]. The model input is a log-mel spectrogram with a shape of $T \times F \times 1$, where T and F denote the number of time frames and frequency bins, respectively. The model consists of two SE blocks and one Transformer encoder, and each SE block consists of two convolutional layers with the kernel sizes of 3×3. The number of output channels of the first and second SE blocks is 64 and 128, respectively. An average pooling layer is applied after each SE block with kernel sizes of 2×2. The Transformer encoder consists of one layer with eight attention heads and a feed-forward network with a hidden size of 32. The final output is obtained by applying max pooling across the time frames T', followed by a fully connected layer.

979-8-3315-7206-8/25/$31.00

TABLE I
SHAPE TRANSITION OF THE BASELINE SE-TRANS MODEL, (FRAMES × FREQUENCIES × CHANNELS), OMITTING SIZE-1 DIMENSIONS.

Module	Input Shape
BatchNorm (bn0)	$T \times F \times 1$
SE Block 1 (pool=2,2)	$(T/2) \times (F/2) \times 64$
SE Block 2 (pool=2,2)	$(T/4) \times (F/2) \times 128$
Adaptive AvgPooling[a] ($T' = 16$)	$T' \times 128$
Transformer Encoder	$T' \times 128$
Temporal Max Pooling	128
Fully Connected Layer	C

[a] nn.AdaptiveAvgPool2d module in PyTorch.

III. PROPOSED APPROACH

A. City-Disjoint Cross-Validation Scheme

In this challenge, evaluating the generalization performance of the model is also important. According to the task requirements, the model must be assessed under domain shift, particularly with respect to unseen cities. The development dataset [4] consists of data from eight cities, while the evaluation dataset is constructed from 12 cities, including six seen cities that overlap with the development dataset and six unseen cities that are not included in the development dataset. This design enables a more comprehensive evaluation under domain shift.

To develop and validate our approach, we adopt a city-disjoint cross-validation scheme. In this scheme, we split the labeled development dataset into two parts, referred to as split 1 and split 2 (see Fig. 1). In the 2-fold cross-validation, we train fold 1 on split 1 and test it on split 2, while fold 2 reverses the roles. The folds are designed to ensure that the cities used for training and testing are disjoint, while the training and validation sets consist of data from the same cities. Note that since the labels "Square" and "Street" were only available in the city "Xi'an", they are not included in this evaluation.

B. Feature-Level Methods

Our proposed approaches are based on two feature-level methods: low-pass filtering applied to the log-mel spectrogram and pitch-shift augmentation applied to the waveform. The overall processing flow is summarized in Fig. 2. These methods are compared in the evaluation experiments in Sec. IV.

1) Low-Pass Filtering: The labeled dataset is limited in size, and as described in the previous section, we further split it, resulting in an even smaller amount of labeled data. This scarcity of data increases the risk of overfitting. To mitigate this, we aim to reduce the complexity of the input data. In this study, we restrict the input features to the low-frequency range, where much of the essential information for ASC is expected to reside. For example, it has been suggested [11] that discriminative content for certain scene classes resides in the low-frequency band. This restriction suppresses redundancy and is expected to mitigate overfitting.

For dimensionality reduction, we adopt a slicing method along the frequency axis, where the dimension F is reduced to a smaller value $\tilde{F}$, resulting in a shape of $T \times \tilde{F} \times 1$. This approach has the advantage that the parameters of the pretrained model can be directly used as initialization during fine-tuning. The only parameters that depend on the frequency dimension are those of the first batch normalization layer (bn0 in Table I). By slicing the input features to match the reduced frequency dimension, the learned statistics in the first batch normalization layer can also be aligned with this dimension, allowing them to be reused effectively.

2) Pitch-Shift Augmentation: As another approach to mitigating overfitting and enhancing robustness to unseen recording conditions, we apply pitch-shift augmentation, which has also been used in previous studies, e.g., [12]. To augment the training set and simulate variations in recording conditions and devices, pitch shifting is applied individually to each sample during training. Specifically, each waveform is pitch-shifted with a probability p, where the shift amount is uniformly sampled from $[-k, +k]$ semitones. This augmentation is applied only to the training samples, while validation and test samples remain unchanged. The transformation is implemented using `librosa.effects.pitch_shift` [13] and applied to the waveform prior to computing the log-mel spectrogram features used as the model input.

C. Temporal Average Pooling

We focus on the feature aggregation method applied to the input of the final fully connected layer, which outputs the final predictions. In the baseline system, max pooling, which extracts the maximum value for each feature dimension across time frames, is employed. Since this operation processes each feature dimension independently, it may not sufficiently reflect the overall temporal structure of the frames.

We introduce feature aggregation methods that explicitly take all time frames into account. Specifically, we investigate two approaches: simple averaging and weighted averaging. In the weighted averaging approach, the weights of each frame are calculated through a fully connected layer, and the features are aggregated based on these weights. This design allows information from all time frames to contribute to the final representation. In the experiments, we compare the baseline max pooling with the two pooling methods, simple average pooling and weighted average pooling.

IV. EXPERIMENT

A. Setup

We used the development dataset provided in the APSIPA ASC 2025 Grand Challenge [4]. All recordings were resampled to a sampling rate of 44,100 Hz. We split the training and validation sets at an 8:2 ratio. The short-time Fourier transform (STFT) was computed using a 40-ms Hanning window with a 20-ms hop size. A set of 64 mel-filter banks was then applied to the spectrograms, followed by a logarithmic operation to obtain log-mel spectrograms. Each log-mel spectrogram had a shape of $T \times F = 500 \times 64$.

	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)

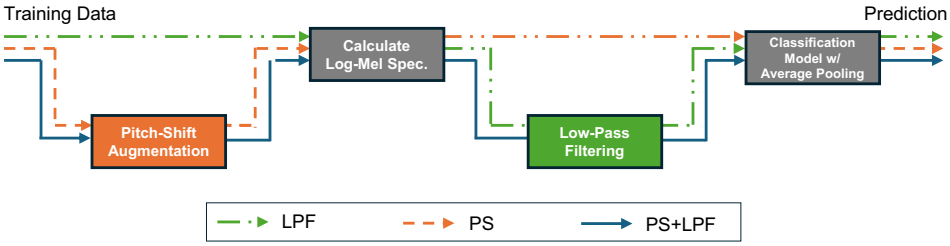

Fig. 2. Processing flow of the proposed approaches. LPF, PS, and PS+LPF denote the approaches using low-pass filtering, pitch-shift augmentation, and their combination, respectively. Note that PS+LPF was not submitted to the challenge but is evaluated in this report for comparison.

In this experiment, we utilized the pre-trained ASC model[1] employed in the baseline system, which had been trained on the TAU Urban Acoustic Scenes (UAS) 2020 Mobile development dataset [3]. We then fine-tuned this pre-trained ASC model using the labeled data from the development dataset. For fine-tuning the baseline model, we used the Adam optimizer with a learning rate of 0.0001 and a batch size of 32. The evaluation metric for this challenge is macro-average accuracy, which is commonly used in previous ASC challenges [2], [3]. This metric is calculated as the average of the class-wise accuracies across the two folds. It should be noted that the test data used in our experiments differs from the official evaluation dataset provided in the challenge.

B. Results

1) Effectiveness of Temporal Average Pooling: In this experiment, we evaluated the effectiveness of temporal aggregation by comparing max pooling (baseline) with average pooling and weighted average pooling. Tables II and III present the comparison results averaged over two folds, corresponding to evaluations on the same cities as the training data and on different cities. The results show that, compared to the baseline, the proposed methods improved the average accuracy by 5 points on the same cities and by 3 points on the different cities. These findings confirmed the effectiveness of averaging across the temporal dimension. However, the performance difference between average pooling and weighted average pooling was marginal. Therefore, we employ average pooling in the subsequent ablation studies on low-pass filtering (i.e., varying $\tilde{F}$) and pitch-shift augmentation (i.e., varying k and p), as it is simple and does not involve any additional learnable parameters.

[1]https://github.com/JishengBai/APSIPA2025GC-ASC/blob/main/pretrained/SE-Trans_pretrained_model.pth

TABLE II
COMPARISON OF ACCURACY FOR EACH LABEL ON THE VALIDATION SET (SAME CITY)

	Max Pool.	Avg. Pool.	Weighted Avg. Pool.
Bus	87.5%	94.8%	92.7%
Airport	93.8%	91.7%	89.6%
Metro	94.7%	94.7%	94.7%
Resto	89.0%	96.2%	93.8%
Mall	87.5%	91.7%	91.7%
Park	82.8%	100.0%	100.0%
Site	87.8%	97.2%	97.2%
Bar	88.6%	100.0%	94.3%
Avg.	89.0%	**95.8%**	94.2%

TABLE III
COMPARISON OF ACCURACY FOR EACH LABEL ON THE TEST SET (DIFFERENT CITY)

	Max Pool.	Avg. Pool.	Weighted Avg. Pool.
Bus	69.6%	70.5%	73.1%
Airport	52.5%	46.9%	45.9%
Metro	64.9%	71.4%	68.9%
Resto	1.0%	3.5%	3.2%
Mall	54.5%	49.6%	46.2%
Park	28.3%	29.6%	33.1%
Site	17.5%	33.5%	36.2%
Bar	19.6%	28.2%	31.5%
Avg.	38.5%	41.6%	**42.3%**

2) Effect of Low-Pass Filtering: In this experiment, we conducted an ablation study on the input feature dimensions, specifically varying $\tilde{F}$, the parameter that controls the cutoff frequency in low-pass filtering (see Sec. III-B1). Table IV shows the results of the evaluation of different cutoff bands for low-pass filtering. The number of frequency bins $\tilde{F}$ was

TABLE IV
EVALUATION OF DIFFERENT CUTOFF BANDS FOR LOW-PASS FILTERING

$\tilde{F}$	macro Acc. (Same)	macro Acc. (Diff.)	Ave.
8	81.7%	37.9%	59.8%
16	86.8%	41.7%	64.2%
24	90.8%	**46.9%**	**68.8%**
32	88.3%	44.4%	66.4%
40	92.7%	43.5%	68.1%
48	93.8%	43.4%	68.6%
56	92.7%	40.2%	66.5%
64	**95.8%**	41.6%	68.7%

TABLE V
EVALUATION OF PITCH-SHIFT PARAMETERS: SEMITONE RANGE k AND PROBABILITY p.

k	p	macro Acc. (Same)	macro Acc. (Diff.)	Ave.
1	0.1	94.8%	41.0%	67.9%
	0.5	94.9%	40.7%	67.8%
2	0.1	**95.1%**	41.1%	**68.1%**
	0.5	94.4%	40.4%	67.4%
5	0.1	93.2%	42.1%	67.7%
	0.5	90.1%	40.2%	65.2%
10	0.1	93.6%	**42.6%**	68.1%
	0.5	89.4%	41.9%	65.6%

set from 8 to 64 in increments of 8. Here, $\tilde{F} = 64$ corresponds to using the full frequency range and is identical to the result of "Avg. Pool." in Tables II and III. On the same cities, the highest average accuracy was achieved with $\tilde{F} = 64$, whereas on the different cities, the highest average accuracy was obtained with $\tilde{F} = 24$. These results suggest that limiting the number of frequency bins can improve classification performance for unseen cities. Among all settings, the highest mean of the two average accuracies was observed with $\tilde{F} = 24$.

3) Effect of Pitch-Shift Augmentation: In this experiment, we conducted an ablation study on pitch shifting, varying the semitone range k and the probability p (see Sec. III-B2). Table V shows the evaluation results for different semitone ranges and probabilities. The semitone range k was set to 1, 2, 5, and 10. For each semitone range, the probability was set to 0.1 and 0.5. On the same cities, the highest average accuracy was achieved with $k = 2, p = 0.1$, whereas on the different cities, the highest average accuracy was obtained with $k = 10, p = 0.1$. Among all settings, the highest mean of the two average accuracies was observed with $k = 2, p = 0.1$.

4) Overall Evaluation of the Proposed Approach: Finally, we compared the best results obtained under each condition: the baseline, low-pass filtering (LPF; reported in Sec. IV-B2), pitch-shift augmentation (PS; reported in Sec. IV-B3), and their combination. Notably, the systems we submitted to the APSIPA ASC 2025 Grand Challenge, named *Kawamura_TMU* and *Masayuki-sera-TMU*[2], correspond to LPF and PS, respectively. Due to time limitations, we could not submit the combination to the Grand Challenge and instead include the results here for completeness. The results are presented in Table VI. When comparing LPF and PS, PS yielded higher accuracy in the same-city evaluation, whereas LPF outperformed in the different-city evaluation. The combination of PS and LPF achieved the best overall performance.

[2]The parameters used for *Masayuki-sera-TMU* in the submission ($k = 2$, $p = 0.5$) differed from those yielding the best ablation result.

TABLE VI
OVERALL EVALUATION. HERE, LPF DENOTES THE RESULTS OF LOW-PASS FILTERING WITH $\tilde{F} = 24$, AND PS DENOTES THE RESULTS OF PITCH-SHIFT AUGMENTATION WITH $k = 2, p = 0.1$.

Method	macro Acc. (Same)	macro Acc. (Diff.)	Ave.
Baseline	89.0%	38.5%	63.7%
LPF	90.8%	46.9%	68.8%
PS	**95.1%**	41.1%	68.1%
PS+LPF	91.5%	**47.3%**	**69.4%**

V. CONCLUSIONS

In this report, we described our submitted system for the APSIPA ASC 2025 Grand Challenge. To evaluate generalization across cities, we adopted a city-disjoint cross-validation scheme based on the provided metadata of city information. We split the labeled development dataset into two folds with non-overlapping training and testing cities. To reduce the risk of overfitting with limited labeled data, we apply two feature-level methods: restricting the input to low-frequency bands to reduce the input complexity and applying pitch-shift augmentation to increase variability. In addition, we replaced max pooling with average pooling to explicitly aggregate information across all time frames. Experimental results demonstrated that the proposed approach consistently outperformed the baseline, achieving approximately an 8-point improvement in accuracy on unseen city environments. These findings suggest that applying feature-level methods and temporal aggregation strategies effectively enhances robustness to unseen cities.

ACKNOWLEDGMENT

This work was supported by JST SICORP Grant Number JPMJSC2306 and JSPS KAKENHI Grant Number JP24KJ1866.

REFERENCES

[1] B. Ding, T. Zhang, C. Wang, *et al.*, "Acoustic scene classification: A comprehensive survey," *Expert Systems with Applications*, vol. 238, p. 121 902, 2024.

[2] A. Mesaros, T. Heittola, E. Benetos, *et al.*, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018.

[3] 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].

[4] 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].

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," in *Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2018, pp. 19–23.

[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.

[11] S. S. R. Phaye, E. Benetos, and Y. Wang, "SubSpectralNet – using sub-spectrogram based convolutional neural networks for acoustic scene classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 825–829.

[12] H. Hu, C.-H. H. Yang, X. Xia, *et al.*, "A two-stage approach to device-robust acoustic scene classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 845–849.

[13] 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.