Evaluation of Low-Frequency Feature Restriction and Average Pooling for Acoustic Scene Classification under Unseen-City Conditions

Takao Kawamura*, Masayuki Sera*, and Nobutaka Ono*
* Tokyo Metropolitan University, Japan

***Abstract*—In this report, we describe our submitted system 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 labeled development set into two folds with non-overlapping training and testing cities, based on the provided metadata of city information. To reduce the risk of overfitting with limited labeled data, we restrict log-mel spectrogram inputs to low-frequency bands, where much of the scene-discriminative information is expected to reside. 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 low-frequency restriction achieves about an 8-point improvement in macro-average accuracy over the baseline in the best configuration.**

I. Introduction

Acoustic Scene Classification (ASC) is the task of identifying acoustic scenes characterized by their surrounding sound environments, such as streets, squares, and restaurants. 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].

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 [4]–[7], 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 [8] 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 set 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 reduce the complexity of the input by using only the low-frequency range, where much of the informative content (e.g., speech in restaurants or announcements in airports) is expected to reside. Third, we replace max pooling in the baseline with average pooling to explicitly aggregate information from all frames. In evaluation experiments, we confirm that our approach achieves an 8-point improvement over the baseline on 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) [9]. The network architecture of the baseline is shown in Table I. In the baseline of the APSIPA grand challenge, the baseline adopts the best configuration of the SE-Trans from [9]. 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 same channels and kernel sizes of 3 × 3. The number of 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. For the Transformer encoder, the baseline sets eight as the number of heads, one as the number of layers, and 32 as the number of units of fully connected layers. The final output is obtained by applying max aggregation across the time frames T', followed by a fully connected layer.

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 different cities. The development dataset [10] 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

TABLE I
SHAPE TRANSITION OF THE BASELINE SE-TRANS MODEL, WHERE SHAPES ARE REPRESENTED AS (FRAMES × FREQUENCIES × CHANNELS) AND DIMENSIONS OF SIZE 1 ARE OMITTED.

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.

unseen cities that are not used during training. 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 set into two folds. 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. The dataset split is summarized in Fig. 1. Here, note that since the labels "Square" and "Street" were only available in the city "Xi'an", they are not included in this evaluation.

B. 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, much of the content in sounds such as conversations in restaurants or announcements in airports is believed to be concentrated 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.

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.

In this report, 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 learnable weighted averaging.

IV. EXPERIMENT

A. Setup

We used the development dataset provided in the APSIPA Grand Challenge [9], [10]. All recordings were resampled to a sampling rate of 44,100 Hz. 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 $T \times F = 500 \times 64$. 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: one evaluated on the same cities as used for training and the other on different cities. The results show that, compared to the baseline, the proposed methods improve the average accuracy by 5 points on the same cities and by 3 points on the different cities. These findings confirm the effectiveness of averaging across the temporal dimension. However, the difference between average pooling and weighted average pooling was small. Therefore, we employ average pooling in the subsequent ablation study on low-pass filtering (i.e., varying $\tilde{F}$), which is simple and does not necessitate additional learned parameters.

2) Effect of Low-Pass Filtering on Model Performance: 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 (described in Sec. III-B). Table IV shows the results of the evaluation of different cutoff bands for low-pass filtering. The number of frequency bins $\tilde{F}$ was set from 8 to 64 in increments of 8. Here, $\tilde{F} = 64$ corresponds to using the full frequency

	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 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%**

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 $\tilde{F} = 24$. Although "Square" and "Street" could not be evaluated, we trained the average pooling model with $\tilde{F} = 24$ using the labeled data in the development dataset and submitted this model.

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 set into two folds with non-overlapping training and testing cities. To reduce the risk of overfitting, we utilized only the low-frequency bands to reduce the complexity of the model input. 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 test environments. These findings suggest that reducing the complexity of the model input and effectively aggregating temporal information are effective strategies for enhancing the robustness to unseen cities.

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%

ACKNOWLEDGMENT

This work was supported by JST SICORP Grant Number JPMJSC2306.

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. DOI: 10.1016/j.eswa.2023.121902.

[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. DOI: 10.1109/TASLP.2017.2778423.

[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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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].