Relaxed Instance Frequency-wise Normalization for Semi-supervised Acoustic Scene Classification

Akansha Tyagi
School of Computing and Electrical Engineering
Indian Institute of Technology, Mandi
E-mail: d19030@students.iitmandi.ac.in

***Abstract*—We present a domain-robust acoustic scene classification system for the APSIPA ASC 2025 challenge that addresses city and temporal variations through Relaxed Instance Frequency-wise Normalization (RFN) and multimodal fusion. Our approach enhances a transformer-based architecture with SE blocks by incorporating RFN to eliminate location-specific and time-dependent domain discrepancies in audio spectrograms. Additionally, we leverage multimodal information by fusing acoustic features with city embeddings and temporal encodings. The system employs a two-stage training strategy with pseudo-labeling on unlabeled data to further improve generalization.**

I. Introduction

Automating a machine to recognize the environmental/surrounding sounds termed as ‘acoustic scenes’ such as a park, market, etc is called acoustic scene classification (ASC) [1]. It is a key part of Computational Auditory Scene Analysis (CASA), has been studied for over ten years and useful in applications where ‘making sense of sound’ is desired [2], like acoustic monitoring, mobile devices with context awareness, smart wheel chairs, etc. Majority of the ASC systems, including the submissions in Detection and Classification of Acoustic Scenes and Events (DCASE) challenge, focus on supervised learning. In contrast, the Asia Pacific Signal and Information Processing Association Annual Summit and Conferene (APSIPA ASC) 2025 challenge takes a different direction. It promotes a semi-supervised learning setup, which better reflects real-world conditions where large amounts of unlabeled data exist alongside only a few labeled samples.

Practical ASC systems face challenges due to different sources of variation, including different recording conditions [3]. These include differences in where the audio is recorded (recording location) and the kind of device used to capture it (recording device). Location-based variation can reflect cultural and regional factors, such as spoken language [4], [5], especially in scenes that involve speech component, like those in the vehicle category. Device-based variation comes from differences in hardware, as different devices have different frequency responses [6]. Such factors lead to intra-scene variation i.e differences between recordings from the same acoustic scene category. In DCASE datasets, recording location and device are key sources of intra-scene variation. While DCASE datasets mainly focus on variations from location and device differences, time-related variation also plays an important role. The Chinese Acoustic Scene (CAS) 2023 dataset highlights this by including timestamp information, allowing analysis of how acoustic scenes change over time. For example, the soundscape of a public square may differ between weekday mornings and weekend evenings. These temporal shifts can affect the background noise, activity level, and presence of specific sound events. ASC systems that ignore these contextual factors struggle to generalize across such variations.

In this work, we propose an enhanced acoustic scene classification framework that addresses domain shift across multiple cities and temporal conditions. Our approach builds upon a transformer-based architecture featuring Squeeze-and-Excitation (SE) blocks for channel-wise attention and transformer encoders for temporal modeling. While this baseline demonstrates strong performance, it remains vulnerable to location-specific and time-dependent variations that degrade accuracy when deployed in unseen cities or different temporal contexts. To address this limitation, we integrate Relaxed Instance Frequency-wise Normalization (RFN) [7], a technique that eliminates instance-specific domain discrepancies by normalizing along the frequency axis where environment-relevant information predominantly resides in audio spectrograms. Unlike conventional channel-based normalization methods, RFN specifically targets the frequency statistics that encode city-specific acoustic characteristics and temporal variations in soundscapes. By incorporating RFN as a plug-and-play module after the initial batch normalization layer, our framework achieves robust generalization across diverse urban environments and time periods while preserving discriminative information essential for accurate scene classification. This integration enables consistent performance across multiple cities and temporal conditions without requiring location-specific or time-aware training procedures, making it particularly suitable for the APSIPA ASC 2025 challenge’s emphasis on cross-city and temporal generalization.

II. Proposed Method

A. Architecture Overview

The system builds upon a transformer-based architecture featuring: (1) Squeeze-and-Excitation (SE) blocks for channel-wise attention, (2) transformer encoders for temporal sequence modeling, and (3) multimodal fusion for incorporating city and temporal metadata. The input audio is first converted to log-mel spectrograms with 64 mel bands, extracted using a 2048-

point FFT with 50% overlap. The spectrograms are processed through two SE blocks with 64 and 128 channels respectively, each followed by 2×2 average pooling. The features then pass through a transformer encoder with 8 attention heads before global pooling and classification.

B. Relaxed Frequency Normalization (RFN)

RFN operates on the principle that domain-relevant information in audio features is dominated by frequency statistics rather than channel statistics. The normalization is formulated as: $\mathrm{RFN}(x) = \lambda \cdot \mathrm{LayerNorm}(x) + (1-\lambda) \cdot \mathrm{FreqInstanceNorm}(x)$, where FreqInstanceNorm normalizes along the frequency axis over batch and time dimensions, while LayerNorm provides global normalization. The relaxation parameter λ (set to 0.5) balances between removing domain-specific artifacts and preserving discriminative information. This mechanism effectively eliminates city-specific frequency patterns while maintaining scene-relevant acoustic features.

C. Multimodal Fusion

To leverage available metadata, we incorporate city and temporal information through learned embeddings:

City Encoding : Each of the 22 cities is mapped to a 16-dimensional embedding vector through a learnable embedding layer, capturing city-specific acoustic priors.

Temporal Encoding : Time metadata is encoded using sinusoidal positional encodings for hour, month, weekday, and minute components, creating an 8-dimensional feature vector that captures cyclical temporal patterns. These features are mapped to a 16-dimensional space through a learned linear transformation. The multimodal fusion occurs after the transformer encoder, where acoustic features (128-dim) are concatenated with city embeddings (16-dim) and temporal embeddings (16-dim), then projected back to 128 dimensions through a fusion layer before final classification.

D. Training Strategy

A two-stage training approach is used with pseudo-labeling:

Stage 1 : The model is trained on labeled data (80% train, 20% validation split) using cross-entropy loss, Adam optimizer with learning rate 1e-4, and step-wise learning rate decay. We initialize from a pre-trained SE-Trans model and fine-tune all layers.

Stage 2 : The trained model generates pseudo-labels for unlabeled samples in the training set. The model is then retrained on the combined labeled and pseudo-labeled data, improving its ability to generalize across domains.

III. Experimental Setup

A. Dataset Description

For the APSIPA ASC 2025 Grand Challenge, the development dataset is a subset of the Chinese Acoustic Scene (CAS) 2023 dataset and contains 24 hours of audio recordings. The data was collected from 22 different cities across China and consists of 10-second audio clips labeled with one of 10 acoustic scene categories namely : 'Bus', 'Airport', 'Metro', 'Restaurant', 'Shopping Mall', 'Public Square', 'Urban Park', 'Traffic Street', 'Construction Site', and 'Bar'.

B. Implementation Details

1) **Feature Extraction :** 64-dimensional log-mel spectrograms with 500 frames
2) **Model Configuration :** 8 attention heads, 32 feed-forward dimensions, 1 transformer layer, dropout rate 0.1
3) **Training :** Batch size 4, 20 epochs maximum with early stopping (patience=10)

References

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, 2015.

[2] T. Virtanen, M. D. Plumbley, and D. Ellis, "Computational analysis of sound scenes and events," *Springer Publishing Company, Incorporated*, vol. 1st, 2017.

[3] P. Jiang, Y. Yang, C. Zou, and Q. Wang, "An attention-based time-frequency pyramid pooling strategy in deep convolutional networks for acoustic scene classification," *IEEE Signal Processing Letters*, 2024.

[4] H. L. Bear, T. Heittola, A. Mesaros, E. Benetos, and T. Virtanen, "City classification from multiple real-world sound scenes," *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 11–15, 2019.

[5] D. Heise and H. L. Bear, "Visually exploring multi-purpose audio data," *23rd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2021.

[6] T. Morocutti, F. Schmid, K. Koutini, and G. Widmer, "Device-robust acoustic scene classification via impulse response augmentation," *31st European Signal Processing Conference (EUSIPCO)*, pp. 176–180, 2023.

[7] B. Kim, S. Yang, J. Kim, H. Park, J. Lee, and S. Chang, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," *arXiv preprint arXiv:2206.12513*, 2022.