SRIB SUBMISSION FOR DCASE 2025 CHALLENGE TASK-1: LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION WITH DEVICE INFORMATION

Technical Report

Krishna G, Ravi S, Sujith V, Madhu R K, Aditi D, Abhinandan U, Ramya V, Rajesh Krishna K S

Samsung Research & Development Institute, Bangalore, India
{krishna.g, ravi.siso, sujith.v, madhu.r, aditi.deo, a.udupa11, r.vishwanath, ks.rajesh}@samsung.com

ABSTRACT

This report details our submission for Task 1: Low-Complexity Acoustic Scene Classification with Device Information in the DCASE2025 challenge[1]. Our method builds upon the leading system from the DCASE2023 competition. Specifically, we have explored the CP-Mobile architecture in this work. To improve the generalization across devices, we incorporate several data augmentation strategies, including Freq-Mix-Style, frequency masking, and time rolling. To meet the model complexity requirements of the competition, we have evaluated the model with 16-bit precision. Hence, we have incorporated the mixed precision training to achieve the better performance during inference with 16-bit model. Our results show significant improvements in test accuracy over the baseline, confirming the effectiveness of our approach across all subsets.

Index Terms— CP-Mobile, Acoustic Scene Classification, Freq–Mix-Style, Frequency masking, Time rolling

1. INTRODUCTION

Acoustic Scene Classification (ASC) aims to categorize recordings into one of multiple predefined acoustic scene categories. The field has advanced significantly due to the annual Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge, held since 2016. This report presents our submission for Task 1 of the DCASE 2025 challenge [1], which focuses on classifying urban acoustic scenes using the TAU Urban Acoustic Scenes dataset [2]. Over the years, the task has evolved to address practical challenges such as device variability and computational efficiency, promoting the development of models suitable for deployment on resource-constrained devices.

The DCASE 2025 challenge task-1 emphasizes several key aspects: (1) recording device mismatches, (2) low-complexity constraints, (3) data efficiency, and (4) the development of recording-device-specific models. Additionally, participants are required to train their models using only 25% of the full dataset, with optional use of external resources permitted under controlled conditions [6]. In addition, strict constraints are imposed on model complexity, limiting parameter size to 128 kB and computational load to a maximum of 30 million multiply-accumulate operations (MMACs) per one-second audio input.

Our method draws inspiration from the leading system in the 2023 competition [3]. This work explored lightweight CP-Mobile [3] model to build the ASC systems. To improve robustness against device mismatch, we incorporate various augmentation techniques, including Frequency masking, Time Rolling, and Freq-Mix-Style [4]. Also, we have incorporated the mixed precision training to achieve the better performance during inference with 16-bit model.

2. EXPERIMENTAL SETUP

2.1. Dataset

The DCASE 2025 challenge employed the TAU Urban Acoustic Scenes 2022 Mobile development dataset (TAU22) [2] for developing the ASC systems under task-1. This database consists audio recordings from 12 European cities, capturing 10 different acoustic scenes using 4 real devices. Further, synthetic data for 11 mobile devices was simulated using original audio recordings. The TAU Urban Acoustic Scenes 2022 Mobile development dataset is consists audio recordings by three real devices (A, B, and C) and six simulated devices (S1-S6). The development dataset comprises of 230,350 audio segments of 1 second duration and all audio segments are in a single-channel, 44.1 kHz, 24-bit format. This year's development set reuses the 25% train split and the test split of Task 1 in the DCASE Challenge 2024.

2.2. Feature extraction

In this work we have developed the systems using Mel-spectral features extracted from audios files which are resampled to 32 kHz and processed to Mel spectrograms with 256/384/512 frequency bins. The Short Time Fourier Transformation uses a window size of 96 ms and a hop size of 16 ms. Spectrograms are computed with 4096-point Fast Fourier Transform (FFT). In addition to Mel-spectral features, we explored a diverse set of acoustic feature representations for the DCASE Task 1 challenge. Specifically, we evaluated Mel-Frequency Cepstral Coefficients (MFCC), and Linear Frequency Cepstral Coefficients (LFCC).

2.3. Data augmentations

In order improve the generalization capability and to avoid overfitting, especially with a relatively small dataset of 25% split, various data augmentation techniques have been used.

1. Time-rolling will shift a randomly segmented starting/trailing portion of duration (up to 0.1 seconds) to the other end of the input audio signal, which can simulate temporal variations audio data [5].
2. Based on previous year's challenge top system [5], in this work, we have adopted Spec-Augmentation with fr equency masking up to 48 frequency bins.

Table 1: Results comparison between different ASC systems.

	macro_avg_acc	NMel	Param	Precision	Model Size (Param*precision in Bytes)	MACs
Baseline (DCASE organizers)	50.70%	256	64K	16 (2B)	128KB	29.4M
BC-Resnet-1	49.94%	256	7.4K	16 (2B)	14.8KB	10.5M
CP-Resnet	53.40%	256	54.7K	16 (2B)	109KB	26.7M
CP-Mobile_S2_256	54.95%	256	61.2K	16 (2B)	122KB	27M
CP-Mobile_S2_384	**56.36%**	384	61.2K	16 (2B)	122KB	27M

Table 2: Ablation results comparison for CP-Mobile based ASC systems.

	macro_avg_acc	Nmel	Stride	Param	Model Size (Param*precision in Bytes)	MACs
CP-Mobile_S2_Data_Balanced	54.54%	256	2	61.2K	119KB	27M
CP-Mobile_S2_256	53.40%	256	2	61.2K	119KB	18M
CP-Mobile_S2_384	**56.36%**	384	2	61.2K	119KB	27M
CP-Mobile_S2_512	57.10%	512	2	61.2K	119KB	37M
CP-Mobile_S1_384	57.20%	384	1	61.2K	122KB	275M

3. Further, we have applied the Freq-Mix-Style [4] augmentation in this study. We apply Freq-Mix-Style to a batch with a probability of 70% and with mixing coefficients are drawn from a Beta distribution with α = 0.3.

2.4. Architecture

This work primary explored the CP-Mobile architectures provided as in [3]. CP-Mobile incorporates CPM block that uses Pointwise expansion and Depthwise convolution followed by a Pointwise projection make it computationally efficient alternative for the conventional convolutional layer. Further the CPM block also includes Batch normalization and ReLU activation. Additionally it uses Residual connection with strided average pooling. More details of the architecture can be found in [3]. Implementation of this architecture can be found at `https://github.com/fschmid56/cpjku_dcase23.git`. In addition to CP-Mobile this work also explored BC-Resnet [6], CP-Resnet [3] to develop the light weight ASC systems.

3. EXPERIMENTAL RESULTS

All the system in this study are trained for 2000 epochs, with initial learning rate of 0.01 and warm-up step of 200. Results of our experiments are presented in Table 1, Table 2 and Table 3, which demonstrate performance of different models on the given development test split dataset. From Table 1 and Table 2 it can be observed that CP-Mobile architecture with Feature dimension 384, offer better performance (56.36%) compared to other systems. When it compared to baseline system, BC-Resnet with very low computational complexity shown competitive performance (49.94%). Further different CP-Mobile variants were investigated, it is observed tha perfromance of ASC system tend to improve with the increase in feature dimension. However with larger feature like NMel=512, the complexity of the system is increased. Also, a the CP-Mobile model with stride=1 offered better results comapred to stride=2, how complexity of the model with stride=1 increased exponentially, makes it not suitable for the challenge. Despite the comprehensive exploration of these features, alternative representations did not yield significant improvements on the validation set. Further we studied the CP-Mobile model performance with different features, and the results are reported in Table-3. Among all the features investigated, baseline Mel-filterbank features promising performance on the validation set when used with the $CP - Mobile$ model. All other feature types consistently under-performed in our experiments. Hence we continued to use baseline Mel-Spectral feature in our study.

Table 3: CP-mobile based Acoustic Scene Classification results for different features

Feature	Val_acc	Input Dim	MACs
MFCC	40.94	40	4.599M
MFCC	44.88	256	29.41M
MFCC	45.14	256	18.57M
MFCC	46.18	384	27.86M
LFCC	37.65	40	4.59M
LFCC	40.18	256	29.41M
LFCC	42.01	256	18.57M
LFCC	43.79	512	37.14M
LFCC	42.76	512	37.14M
Mel filterbank (baseline)	51.87	256	29.41M

Table 4: Final Results Comparison.

Device ID	General training		Device-specific Training	
	Baseline	CP-Mobile_S2_384	Baseline	CP-Mobile_S2_384
a	62.80%	**67.48%**	63.98%	**69.33%**
b	52.87%	**57.69%**	55.85%	**59.76%**
c	54.23%	**60.73%**	59.09%	**62.61%**
s1	48.52%	**54.42%**	48.68%	**55.97%**
s2	47.29%	**52.12%**	48.74%	**53.45%**
s3	52.86%	**59.03%**	52.72%	**59.12%**
s4	48.14%	**52.97%**	48.14%	**52.61%**
s5	47.23%	**54.30%**	47.23%	**54.21%**
s6	42.60%	**48.52%**	42.60%	**48.36%**
Average	50.73%	**56.36%**	51.89%	**57.27%**

From Table 4, it can be observed that the propose system $CP - Mobile\_S2\_384$ (stride=2, NMel=384) offered better results on validation set in general training task. Further this system considered for device specific model training. From results it can be observed the proposed $CP - Mobile\_S2\_384$ outperformed the baseline system by 6% in terms of absolute accuracy.

4. CONCLUSION

In this technical report, we present the SRIB submission to Task-1 of the DCASE 25 challenge. We show that our system outperforms the baseline system by DCASE 25 challenge by 6% in terms of accuracy and the our system can match its performance with similar number of parameters and MACs of the baseline system. The reason for this improvement can be attributed to an efficient CP-Mobile architecture. And it is observed that CP-Mobile architecture with higher feature dimension shows better performance. When it compared to baseline system, BC-Resnet with very low computational complexity shown competitive performance.

5. REFERENCES

[1] F. Schmid, P. Primus, T. Heittola, A. Mesaros, I. Martín-Morató, and G. Widmer, "Low-complexity acoustic scene classification with device information in the dcase 2025 challenge," 2025. [Online]. Available: https://arxiv.org/abs/2505.01747

[2] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, pp. 56–60. [Online]. Available: https://arxiv.org/abs/2005.14623

[3] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, "Cp-jku submission to dcase23: Efficient acoustic scene classification with cp-mobile," *Tech. Rep., DCASE2023 Challenge*, 2023.

[4] 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.

[5] D. Nadrchal, A. Rostamza, and P. Schilcher, "Data-efficient acoustic scene classification with pre-trained cp-mobile," DCASE2024 Challenge, Tech. Rep, Tech. Rep., 2024.

[6] B. Kim, S. Chang, J. Lee, and D. Sung, "Broadcasted Residual Learning for Efficient Keyword Spotting," in *Proc. Interspeech 2021*, 2021, pp. 4538–4542.