MCCI SUBMISSION TO DCASE 2025: TRAINING LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION SYSTEM WITH KNOWLEDGE DISTILLATION AND CURRICULUM

Technical Report

*Xuanyan Chen**

Guangxi University
School of Computer, Electronics and Information
Nanning, China
sham@st.gxu.edu.cn

Wei Xie†

Guangxi University
School of Computer, Electronics and Information
Nanning, China
chester.w.xie@gmail.com

ABSTRACT

The Task 1 of DCASE 2025 focuses on different aspects of Acoustic Scene Classification(ASC) including recording device mismatch, low complexity constraints, data efficiency and the development of recording-device-specific models. This technical report describes the system we submitted. We first trained several teacher models on the ASC dataset through Self-Distillation and Curriculum Learning techniques.These teacher models included a model pre-trained on the AudioSet. Then we distill the knowledge from the teacher model into the student model via curriculum learning. We used the same inference model (i.e., student model) and data augmentation settings as provided in the baseline system. In experiments, our best system achieved an accuracy of 57.66%.

Index Terms— Acoustic Scene Classification, knowledge distillation, curriculum learning

1. INTRODUCTION

Acoustic scene classification systems categorize recordings into one of multiple predefined acoustic scene classes. The Task 1 of DCASE 2025 [1] focus on different aspects in ASC including recording device mismatch, low complexity constraints, data efficiency and the development of recording-device-specific models. Compared to last year, Task 1 of DCASE 2025 provides device ID in the evaluation set, and participants can choose to train independent models for different recording devices or develop generic models with strong generalizability, as the evaluation set contains devices that are not present in the training set. Meanwhile, last year's task focused on the problem of data efficiency. The training set was divided into five subsets containing 5%, 10%, 25%, 50%, and 100% of the complete data respectively. The average score of the participating systems on these five subsets was taken. In DCASE 2025, participants were only allowed to train with a subset of 25% of the complete training data.

In recent years' tasks, Knowledge Distillation (KD) has been proven to be an excellent approach for addressing low-complexity constraints and data efficiency issues [2]. Researchers typically first train some high-performance complex models as teacher models, and then transfer the knowledge from the teacher models to student models through knowledge distillation techniques. The student models are usually lightweight models developed to solve low-complexity constraints. Data augmentation techniques such as Freq-Mixup and Device Impulse Response (DIR) Augmentation have been applied to tackle problems of recording device mismatch and generalization. We also found that Curriculum Learning(CL) [3], as a training strategy, can help models utilize data more effectively, improve model generalization, and accelerate the convergence speed [4]. Therefore, in this report, we propose a method that combines knowledge distillation with curriculum learning. First, we trained several teacher models on the training set using self-distillation and curriculum learning techniques, and then transferred the knowledge to student models through knowledge distillation and curriculum learning. Experiments have demonstrated that this approach enhances the classification performance of student model.

2. METHOD

2.1. Self-Distillation

To obtain better-performing teacher models, we designed a self-distillation [5] method for the teacher models. First, the teacher model was divided into 4 blocks, and a feature alignment module and a classifier were added after the first three blocks. These additional modules, together with the model backbone, were optimized during the training process. In the training phase, the last classifier was used as the teacher model, and the first three classifiers were used as student models for knowledge distillation. In the inference phase, the ensemble of outputs from the four classifiers was taken as the final output of the model. Experiments have proven that this method can effectively improve the performance of the models.

2.2. Curriculum Learning

Inspired by [6], we designed a curriculum learning method for knowledge distillation. We assigned unique parameters to all categories and samples in the training set, which replace the temperature in knowledge distillation and are co-optimized with model parameters during the training phase. Specifically, when the model correctly classifies a sample, it tends to be regarded as an easy-to-learn sample, and its corresponding category parameter and sample parameter increase (i.e., the distillation temperature increases). Therefore, the contribution of this sample to gradient update is enhanced, the model pays more attention to these samples, and learns more "dark knowledge" from them, and vice versa.

*
†

Table 1: Configuration and performance evaluation on test set of three submission systems.

System ID	Parameter Num	MMACs	Teacher Models	ACC
S1	61148	29419156	5 CNN + 2 Transformer	56.54
S2	61148	29419156	5 CNN	57.66
S3	61148	29419156	2 Transformer	56.06

3. EXPERIMENTAL SETUP

3.1. DataSet

As required by the task, we used a subset of the TAU Urban Acoustic Scenes 2022 Mobile development dataset, which contain 25% complete training data. The dataset contains recordings from 12 European cities in 10 different acoustic scenes using 4 devices. Additionally, synthetic data for 11 mobile devices was created based on the original recordings. Of the 12 cities, only two are present in the evaluation set. The dataset has exactly the same content as the TAU Urban Acoustic Scenes 2020 Mobile development dataset, but the audio files have a length of 1 second (therefore, there are ten times more files than in the 2020 version). In summary, its training and test sets contain 34,900 and 29,680 audio respectively.

Recordings were made using four devices that captured audio simultaneously. The primary recording device, referred to as device A, consists of a Soundman OKM II Klassik/studio A3, an electret binaural microphone, and a Zoom F8 audio recorder using a 48kHz sampling rate and 24-bit resolution. The other devices are commonly available consumer devices: device B is a Samsung Galaxy S7, device C is an iPhone SE, and device D is a GoPro Hero5 Session.

3.2. Teacher Model

In reference to [7], we selected CP-ResNet [8], CP-Mobile [9], and PaSST [10] as teacher models, where CP-ResNet and CP-Mobile are variants of CNN, and PaSST is a variant of Transformer. We combined CP-ResNet and CP-Mobile with the self-distillation and curriculum learning methods mentioned earlier for training, resulting in a total of 5 different teacher models. For PaSST, we chose a randomly initialized model and a model pre-trained on AudioSet [11] as teacher models. The overall situation of the teacher models is shown in Table 2.

3.3. Training Settings

For training the model, audio input is resampled to 32 kHz and converted to mel spectrograms using a 4096-point FFT with a window size of 96 ms and a hop size of approximately 16 ms, followed by a mel transformation with a filterbank of 256 mel bins. The system is trained for 200 epochs using the SGD optimizer and a batch size of 256. Freq-MixStyle is applied to tackle the device mismatch problem, and time rolling of the waveform and frequency masking are used to augment the training data. The base line system requires 29.4 MMACs for the inference on a one-second audio clip. The memory required for the model parameters amounts to 122.3 kB, resulting from the 61,148 parameters used in 16-bit precision (float 16).

Table 2: Overall situation of the teacher models.

Type	Method	Acc
CNN	CP-ResNet	54.16
	CP-ResNet + sd	55.34
	CP-ResNet + sd +cur	56.24
	CP-Mobile + sd	55.89
	CP-Mobile + sd + cur	56.43
Transformer	PaSST	43.52
	PaSST-pt	53.86

4. SUBMISSION AND RESULT

The configuration differences and the performance on test set of the three systems we submitted are shown in Table 1. The best system achieved an accuracy of 57.66% on the test set.

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] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, "Distilling the knowledge of transformers and CNNs with CP-mobile," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, 2023, pp. 161–165.

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[4] X. Wang, Y. Chen, and W. Zhu, "A survey on curriculum learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 4555–4576, 2021.

[5] L. Zhang, C. Bao, and K. Ma, "Self-distillation: Towards efficient and compact neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4388–4403, 2021.

[6] S. Saxena, O. Tuzel, and D. DeCoste, "Data parameters: A new family of parameters for learning a differentiable curriculum," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[7] B. Han, W. Huang, Z. Chen, A. Jiang, P. Fan, C. Lu, Z. Lv, J. Liu, W.-Q. Zhang, and Y. Qian, "Data-efficient low-complexity acoustic scene classification via distilling and progressive pruning," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[8] B. Kim, S. Yang, J. Kim, and S. Chang, "Qti submission to dcase 2021: Residual normalization for device-imbalanced acoustic scene classification with efficient design," *arXiv preprint arXiv:2206.13909*, 2022.

[9] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Receptive field regularization techniques for audio classification and tagging with deep convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1987–2000, 2021.

[10] K. Koutini, J. Schlüter, H. Eghbal-Zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021.

[11] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.