GISP@HEU'S SUBMISSION FOR DCASE 2026 TASK 7: PROTOTYPE-GUIDED EXPERT NETWORK FOR DOMAIN-INCREMENTAL LEARNING

Technical Report

Xuefeng Yang[1], *Tong Ye*[1], *Xiaoyu Feng*[1], *Wenbo Wang*[1], *Qiaoxi Zhu*[2], *Jian Guan*[1*]

[1]Group of Intelligent Signal Processing, Harbin Engineering University, Harbin, China
[2]University of Technology Sydney, Ultimo, Australia

ABSTRACT

This report presents our submission to DCASE 2026 Task 7 on domain-agnostic incremental audio classification, where models learn D1→D2→D3 without revisiting previous-domain data and automatically infer test domains. All systems we submitted use a domain-incremental backbone with a frozen shared encoder, domain-specific BatchNorm, incremental experts, and a prototype-gated selector. System 1 adds a spectral-temporal dual-branch head for reverberant D3. System 2 evaluates weighted prediction error (WPE) dereverberation on D3. System 3 combines raw, WPE, and multi-window dereverberated predictions in a three-way ensemble with conservative routing. System 4 uses the same ensemble but lowers the fallback threshold to favor D2 and D3 routing. On the development set, all systems achieve 65.31%–68.18% average accuracy across D2 and D3, surpassing the 52.50% official baseline. System 4 performs best, reaching 68.18% average accuracy, including 72.78% on D2 and 63.59% on D3.

Index Terms— Domain-incremental learning, acoustic scene classification, mixture of experts

1. INTRODUCTION

DCASE 2026 Challenge Task 7 aims to develop a universal domain-incremental learning (DIL) system for audio classification [1]. In this task, the system is required to sequentially adapt to three recording domains in a fixed order, i.e., $D_1 \rightarrow D_2 \rightarrow D_3$. Each domain contains the same 10 sound-event classes (*alarm, baby, dog, engine, fire, footsteps, knock, phone, piano, speech*), while only the data from the current domain is available at each training stage [1, 2]. This setting makes the task particularly challenging in two aspects. First, since previous-domain data cannot be revisited when adapting to a new domain, the model must retain previously acquired knowledge and avoid catastrophic forgetting. Second, the evaluation is domain-agnostic, meaning that only the audio signal is provided at test time and the system must infer the domain of each clip before applying the corresponding classifier.

To address these challenges, we develop four domain-incremental systems that improve domain adaptation, reverberation robustness, and domain-agnostic routing from different perspectives. System 1 combines a domain-incremental backbone [1, 3] with a Gaussian prototype-gated selector [4, 5], and introduces a spectral-temporal dual-branch pooling head to enhance representation learning for the reverberant D_3 domain. System 2 investigates the independent effect of single-channel dereverberation by applying a weighted prediction error (WPE) [6] blind dereverberation front-end to D_3. System 3 integrates the designs of Systems 1 and 2 into a three-way ensemble, where predictions from raw and dereverberated variants are fused under a conservative routing strategy. System 4 adopts the same ensemble backbone as System 3 but uses a lower fallback threshold to favor routing toward D_2/D_3 when sufficient prototype evidence is available. These four systems constitute our submissions to DCASE 2026 Task 7.

**Corresponding author.

2. SUBMISSION SYSTEM

2.1. System 1

System 1 combines a domain-incremental backbone [7] with a prototype-gated selector. The backbone uses a frozen shared encoder, domain-specific BatchNorm layers, and incrementally trained deep experts for D_2 and D_3, preserving prior knowledge while adapting to new domains. For reverberant D_3, a spectral-temporal dual-branch pooling head captures complementary frequency and temporal cues, while D_1 and D_2 use standard global pooling. During inference, Gaussian prototype matching routes [3] uncertain samples to D_1 and assigns others to the higher-scoring domain between D_2 and D_3.

2.2. System 2

System 2 focuses on evaluating the contribution of single-channel dereverberation independently. Building on the same domain-incremental backbone as System 1, it extends the pipeline with only a WPE blind dereverberation front-end applied to the reverberant domain D3, while D1 and D2 keep their original audio. Notably, during both the incremental training of D3 and the inference stage, the training and test clips used in this process are first dereverberated and then fed into the shallow encoder.

2.3. System 3

System 3 combines the key designs of Systems 1 and 2 in a three-way ensemble with conservative domain routing. It adopts the same frozen shallow encoder with shared convolutions and domain-specific BatchNorm as Systems 1 and 2. As for the incrementally trained deep experts, D1 and D2 use standard global-pooling experts. For reverberant D3, softmax probabilities from the raw signal, the WPE-dereverberated signal, and a multi-window dereverberated variant [8] are averaged. Finally, domain routing uses the prototype-gated selector with a fallback threshold of $Q = 30\%$: un-

Table 1: Development-set results under the official domain-agnostic protocol (overall accuracy, %).

Method	D2	D3	Avg.
Official baseline	58.60	46.10	52.50
System 1	73.47	60.03	66.75
System 2	71.74	61.34	66.54
System 3	69.29	61.33	65.31
System 4	72.78	63.59	**68.18**

certain clips are assigned to D1, and the rest to the higher-scoring domain between D2 and D3.

2.4. System 4

System 4 shares System 3's three-way ensemble: a frozen shallow encoder, domain-specific BatchNorm, and incremental deep experts [3] for D2 and D3. The only change is routing: a prototype-gated selector with a low fallback threshold ($Q = 1\%$) sends only clips clearly dissimilar to D2/D3 to D1, assigning all others to the higher-scoring domain.

3. EXPERIMENTAL RESULTS

3.1. Dataset

We use the DIL−DCASE26 dataset from DCASE 2026 Task 7 [1], which focuses on domain-agnostic incremental audio classification. The dataset contains 10 sound classes: *alarm, baby_cry, bark, engine, fire, footsteps, knock, telephone_ringing, piano*, and *speech*. The development set includes two domains, D2 and D3, while knowledge of D1 is provided through the official baseline model.

3.2. Experimental Setup

We formulate DCASE 2026 Task 7 as domain-incremental audio classification over D1–D3 with ten machine-sound classes. All systems use 32 kHz mono audio, 64-bin log-mel features following PANNs preprocessing [9], 4-second training segments, and 4s/2s sliding-window averaging at inference. Starting from the official D1 CNN14 checkpoint [1], we freeze shallow convolutional layers and sequentially adapt domain-specific BatchNorm layers and deep experts for D2 and D3. Remix [10] and SpecAugment [11] address class imbalance and robustness. At inference, all systems use prototype-guided domain selection, differing only in D3 processing: System 1 adds a dual-branch head, System 2 applies WPE dereverberation. Systems 3 and 4 ensemble raw, dereverberated, and multi-window routes with different fallback thresholds.

Here, we only report the average accuracy on D2 and D3 in our experiments, and the final challenge ranking is based on the overall average accuracy across D1, D2, and D3.

3.3. Results

Our four systems obtain the average macro accuracy between 65% and 68% on D2 and D3, all exceeding the official baseline of 52.5%. In which, system 4 achieves the best overall performance, with the hightest D3 accuracy at 63.59%.

4. CONCLUSION

We presented our DCASE 2026 Task 7 submission for domain-incremental audio classification. All four systems use a shared framework with a frozen shallow encoder, domain-specific experts, and a prototype-gated domain selector to reduce forgetting under a domain-agnostic protocol. They trade off complexity and robustness through different ensemble, dereverberation, test-time augmentation, and fallback strategies. All systems outperform the official baseline, with the best reaching a 68.18% D2/D3 average. Future work will focus on better unknown-D1 detection and robustness in reverberant domains.

5. REFERENCES

[1] R. Casciotti, M. Mulimani, M. Harju, J. R. Jensen, and A. Mesaros, "Domain-agnostic incremental learning for sound classification. a DCASE 2026 challenge task," 2026, arXiv preprint arXiv:2606.02173.

[2] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, 2022.

[3] M. Qin, X. Zhang, X. Wang, K. Wei, X. Yang, and C. Deng, "DIMoE-Adapters: Dynamic expert evolution for continual learning in vision-language models," 2026, arXiv preprint arXiv:2605.07494.

[4] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4077–4087.

[5] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 7167–7177.

[6] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[7] M. Mulimani and A. Mesaros, "Domain-incremental learning for audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[8] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, "Better aggregation in test-time augmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1214–1223.

[9] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2880–2894, 2020.

[10] H.-P. Chou, S.-C. Chang, J.-Y. Pan, W. Wei, and D.-C. Juan, "Remix: Rebalanced mixup," in *Computer Vision – European Conference on Computer Vision (ECCV) 2020 Workshops*, 2020, pp. 95–110.

[11] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. INTERSPEECH*, 2019, pp. 2613–2617.