UNIS SYSTEM FOR NOISE-AWARE UNSUPERVISED ANOMALOUS SOUND DETECTION FOR MACHINE CONDITION MONITORING

Technical Report

Junjie Wang

Independent Researcher
1664171068@qq.com

ABSTRACT

This technical report presents our solution to Task 2 of the DCASE 2026 Challenge. We developed four subsystems for unsupervised anomalous sound detection, all of which extract audio embeddings from fine-tuned audio pre-trained models and perform anomaly detection using outlier detection algorithms. Compared with previous editions of the challenge, the target machine sounds in this year's task were synchronously recorded by multiple microphones, providing both near-field and far-field acoustic views. To exploit this characteristic, we leverage far-field recordings as a data augmentation strategy, enabling the models to learn more robust acoustic representations and improving their generalization ability under different recording conditions and noisy environments. In addition, we employ multiple pre-trained models to obtain complementary feature representations, further enhancing the overall anomaly detection performance.The source code is publicly available at: `https://github.com/outman-goutian/dcase2026_task2`.

***Index Terms*—** Self-supervised, pre-trained, domain generalization

1. INTRODUCTION

This technical report presents our solution to Task 2 of the DCASE 2026 Challenge [1–4], which addresses the problem of unsupervised anomalous sound detection for machine condition monitoring. Compared with previous editions, this year's challenge introduces a multi-microphone recording setting in realistic scenarios, where target machine sounds are captured simultaneously using both near-field and far-field microphones, providing richer acoustic information for anomaly detection.

We adopt a self-supervised learning framework by constructing an auxiliary classification task to learn audio embeddings, which are then used to compute anomaly scores via KMeans clustering. Conventional approaches typically rely on machine metadata as the supervision signal for the auxiliary task. To enhance the discriminative power of the learned embeddings, we introduce audio pretrained models, namely BEATs [5] and EAT [6], instead of training models from scratch, thereby improving feature representation capability.

In addition, we exploit far-field audio as a data augmentation strategy by randomly mixing it with near-field audio, simulating variations in recording distance and acoustic conditions. This design further improves the robustness of the learned embeddings and enhances the generalization ability of the proposed system.

2. PROPOSED ASD SYSTEM

2.1. Proposed Method

This paper proposes an unsupervised anomalous sound detection method based on pretrained audio representation models. We employ EAT-Base and BEATs as the backbone feature extractors and perform full fine-tuning on the DCASE 2026 Task 2 dataset to learn discriminative audio representations tailored to the target task. During training, ArcFace Loss [7] is adopted to optimize the embedding space by encouraging intra-class compactness and inter-class separability. During inference, the anomaly score is computed using the K-Nearest Neighbors (KNN) algorithm based on the distance between the test embedding and the reference embeddings of normal training samples, enabling unsupervised anomaly detection.

2.2. Submitted Systems

We submitted four systems for DCASE 2026 Task 2. The first system uses EAT-Base as the backbone model, while the second system adopts BEATs as the backbone. The remaining two systems are score-level fusion systems. Specifically, the third system combines the anomaly scores of EAT-Base and BEATs with fusion weights of 0.45 and 0.55, respectively, whereas the fourth system uses equal fusion weights of 0.5 and 0.5.

2.3. Result

The experimental results demonstrate that the systems based on pretrained audio representation models consistently outperform conventional baseline methods in the target domain. Compared with the MSE and MAHALA baselines, the proposed System 3 achieves superior or competitive performance across most machine types, with notable improvements on categories such as ToyCar, Gearbox, and Valve. Moreover, System 3 exhibits stable performance after fusing EAT-Base and BEATs, which validates the effectiveness of combining complementary pretrained audio representations.

3. CONCLUSION

We propose an unsupervised anomalous sound detection system for the DCASE 2026 Task 2 challenge. The framework leverages pretrained audio models, including BEATs and EAT, to extract discriminative embeddings, which are further optimized using ArcFace loss. Anomaly detection is performed using a KNN-based distance measure in the learned embedding space. To address the multi-microphone recording scenario, far-field audio is incorporated as

Table 1: Comparison with baseline methods and the proposed system (System 3) on the development set.

Machine Type	MSE			MAHALA			System 3		
	Source AUC	Target AUC	pAUC	Source AUC	Target AUC	pAUC	Source AUC	Target AUC	pAUC
ToyCar (Emu)	69.62	61.20	55.89	69.49	66.62	53.47	60.72	95.44	55.21
ToyCar	75.62	37.87	54.03	77.28	53.17	58.25	68.92	89.72	62.66
Bearing (Emu)	62.34	59.56	59.85	65.92	62.28	60.42	67.64	63.32	61.42
Fan	61.45	46.94	53.33	60.00	45.09	52.29	75.18	50.20	52.18
Gearbox (Emu)	68.23	49.78	52.94	74.48	52.74	53.97	76.18	68.24	54.47
Slider (Emu)	67.25	45.05	50.38	66.36	49.18	50.36	71.20	64.16	52.95
Valve (Emu)	67.74	68.78	55.08	56.60	56.50	50.20	99.44	70.50	78.89

a data augmentation strategy to improve robustness under varying acoustic conditions. In addition, we explore multi-model fusion at both feature and score levels to enhance overall performance. Experimental results demonstrate that the proposed approach consistently outperforms baseline methods across most machine categories, with the fusion system achieving the most stable performance.

4. REFERENCES

[1] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[2] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," *Proceedings of 31st European Signal Processing Conference (EUSIPCO)*, pp. 191–195, 2023.

[3] T. Nishida, N. Harada, D. Takeuchi, D. Niizumi, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2026 challenge task 2: Noise-aware unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2606.01578*, 2026.

[4] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1–5.

[5] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[6] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "Eat: Self-supervised pre-training with efficient audio transformer," 2024.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.