ENHANCED ANOMALY DETECTION APPROACH FOR DCASE 2025 TASK 2

Technical Report

Guirui Zhong[1], *Qing Wang*[1], *Jun Du*[1]

[1] University of Science and Technology of China, Hefei, China
skye@mail.ustc.edu.cn,{qingwang2, jundu}@ustc.edu.cn

ABSTRACT

Addressing the unique challenge of the DCASE 2025 Task 2, where the availability of clean machine and noise-only data varies and datasets in previous years are introduced, we propose an enhanced anomaly detection approach that combines data augmentation and two-stage pre-training methods using pre-trained audio separation and self-supervised learning (SSL) models, respectively. Leveraging audio separation models guided by clean machine or noise-only data, our system can separate clean data from noisy data and generate more diverse data in the training phase. Using a lot of machine sound data for two-stage pre-training, the system can better adapt to anomalous sound detection (ASD) task in the downstream fine-tuning task. By integrating these approaches, our system achieves a better performance across different machines on the DCASE 2025 ASD development dataset, ensuring reliable anomaly detection in machine condition monitoring applications.

Index Terms— Anomalous sound detection, data augmentation, audio separation, two-stage pre-training

1. INTRODUCTION

In DCASE challenge 2025 Task 2 "*First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring*" [1, 2], it is required to detect anomalous sounds of machines. In real-world conditions, it is often easier for us to obtain the sound of the machine working normally, while the anomalies are rare and highly diverse. Therefore, we need to use the normal sounds in the training data to detect anomalous sounds in the test data. Furthermore, the operational states of a machine or the environmental noise can change to cause domain shifts. The system needs to use domain generalization techniques to handle frequent or hard-to-notice domain shifts. In the DCASE 2025 task, consistent with last year, the first-shot problem is still introduced, whose two main features are training a model for a completely new machine type and using a limited number of machines from its machine type when training a model. Meanwhile, there still existing some machine types lack of metadata to describe their operation states in details. However, unlike last year, one new change is introduced in the DCASE 2025 task, that is, we can optionally use additional clean machine data or noise-only data (supplemental data) to train a model, which means that we can use supplemental data to implement audio separation.

Our submission includes two major approaches for anomalous sound detection. To begin with, for those machines with clean machine data, we use SepReformer [3] to realize machine sound separation guided by corresponding clean data. Subsequently, we mix the separated clean machine sounds with the noise in the dataset to obtain noisy synthetic samples for data augmentation. Secondly, we propose a two-stage pre-training method to better adapt to ASD task in the fine-tuning period. Because of the introduction of other years' machine sound data, we can carry out further pre-training process on the basis of first general audio dataset pre-training stage.

Each recording used in this challenge is a single-channel. For the case that the recording duration of part different machine types is inconsistent, we process the duration of all audio to 10 seconds by padding or truncating. The development set includes seven machines: ToyCar, ToyTrain, Fan, Gearbox, Bearing, Slide rail, and Valve, and the evaluation set includes eight new machines: AutoTrash, HomeCamera, ToyPet, ToyRCCar, BandSealer, Polisher, ScrewFeeder, and CoffeeGrinder [4, 5]. In the following, we will describe each approach and our experimental results in detail.

2. PROPOSED APPROACH

2.1. Audio Separation

Because some machines have corresponding clean data, we use machine audio separation guided by clean data to separate clean sounds from noisy sounds. Specifically, we employ SepReformer to realize machine audio separation guided by corresponding clean data. Through audio separation, we get clean sound from noisy sound. Then, we use the clean separated sound for further data augmentation. The separated sound is further synthesized with the noise in the data set to obtain noisy sound, and the number of samples in source and target domains is controlled to be 4:1 to alleviate the problem of domain shift problem. Finally, we get a more diverse dataset with even domain distribution. We combine the synthetic data, separated clean data, and original data to train our model later.

2.2. Two-stage Pre-training

Since the attribute information of some machine types is available, we can train a classifier with machine attribute information. Such anomalous sound detection methods based on self-supervised classification have been used in previous challenges and achieved good results [6, 7, 8, 9, 10]. Moreover, due to the excellent performance of the pre-training method, we fine-tune EAT [11] as our backbone to implement attribute classification task. However, unlike previous works, we combine a two-stage pre-training method to better adapt to the ASD task in the downstream fine-tuning task. Specifically, we use DCASE task 2 data from all years to further pre-train the backbone on the basis of the first AudioSet [12] pre-training stage.

Firstly, we transformed all audio clip into spectrograms with a Mel transformation. Then, the audio feature is split into multiple patches and input into the ViT backbone pre-trained on the AudioSet (first stage) and DCASE task 2 datasets (second stage).

Table 1: DCASE 2025 Task 2 experimental results on development dataset (%). The value in the row "Total Score" represents the harmonic mean of the AUC and pAUC scores over all the machine types, sections, and domains.

		Baseline (AE-MSE)	Baseline (AE-MAHALA)	Our system
ToyCar	AUC (source)	71.05	73.17	**96.82**
	AUC (target)	53.52	50.91	**78.84**
	pAUC	49.70	49.05	**63.57**
ToyTrain	AUC (source)	**61.76**	50.87	57.96
	AUC (target)	56.46	46.15	**62.72**
	pAUC	**50.19**	48.32	49.89
bearing	AUC (source)	**66.53**	63.63	55.80
	AUC (target)	53.15	59.03	**63.84**
	pAUC	61.12	**61.86**	53.68
fan	AUC (source)	70.96	**77.99**	56.44
	AUC (target)	38.75	38.56	**58.60**
	pAUC	49.46	50.82	**52.05**
gearbox	AUC (source)	64.80	73.26	**80.16**
	AUC (target)	50.49	51.61	**76.08**
	pAUC	52.49	55.07	**62.05**
slider	AUC (source)	70.10	73.79	**78.50**
	AUC (target)	48.77	50.27	**62.94**
	pAUC	52.32	53.61	**54.05**
valve	AUC (source)	63.53	56.22	**78.62**
	AUC (target)	67.18	61.00	**86.70**
	pAUC	57.35	52.53	**67.94**
Total Score		56.26	55.33	**64.41**

The backbone models each patch and outputs embeddings for all patches. Thereafter, the attentive statistics pooling layer [13] is used to merge all patches information into an utterance embedding. Lastly, the utterance embedding is mapped into low-dimensional embedding. This embedding is utilized for machine attribute classification by ArcFace [14] classifier during fine-tuning. We combine each machine type, domain, and its corresponding attributes as a separate class and perform classification tasks. For those machine types without attributes, we combine their machine types and domains as separate classes. After fine-tuning, this embedding is used for anomaly detection in the backend during testing.

2.3. Backend

To start with, we train KNN detectors for each machine type, using the embeddings from all training samples. Subsequently, the anomaly score for a given testing machine audio is obtained by computing the cosine distance between the embedding of the testing audio and its closest neighbor (k = 1). Meanwhile, we also compute the anomaly score of the corresponding separated clean sound and add these two anomaly scores as the final anomaly score.

2.4. Implementation

All audio waveforms are padded or truncated to 10s, and then converted to log-mel spectrograms with a frame length of 25ms, a frame shift of 10ms, and 128 mel bins. We use the Adam optimizer as the optimizer. For the second pre-training stage, we pre-train the EAT backbone for 10k iterations with a batch size of 16. The two-stage pre-trained EAT is fine-tuned for 30 epochs with a batch size of 32. A cosine learning rate scheduler is adopted with an upper limit of 5e-5 in two-stage pre-training and fine-tuning periods.

2.5. Submission

In this challenge, we submit four different systems with the difference of the data used in audio separation and two-stage pre-training methods. Specifically, we generate 10k or 20k synthetic data for fine-tuning and use machine sound data of only 2025 years or from all years for two-stage pre-training.

2.6. Results

Table 1 shows our best results on the development set using the combination of two anomaly detection approaches in our four submission systems.

3. CONCLUSIONS

In this paper, we propose an enhanced method for anomalous sound detection based on audio separation and two-stage pre-training methods. Experimental results show that by integrating these approaches we can achieve better results than the baseline under available and unavailable attributes.

4. REFERENCES

[1] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2506.10097*, 2025.

[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] U.-H. Shin, S. Lee, T. Kim, and H.-M. Park, "Separate and reconstruct: Asymmetric encoder-decoder for speech separation," *arXiv preprint arXiv:2406.05983*, 2024.

[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] 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.

[6] R. Giri, S. V. Tenneti, K. Helwani, F. Cheng, U. Isik, and A. Krishnaswamy, "Unsupervised anomalous sound detection

using self-supervised classification and group masked autoencoder for density estimation," DCASE2020 Challenge, Tech. Rep., July 2020.

[7] J. Lopez, G. Stemmer, and P. Lopez-Meyer, "Ensemble of complementary anomaly detectors under domain shifted conditions," DCASE2021 Challenge, Tech. Rep., July 2021.

[8] Y. Zeng, H. Liu, L. Xu, Y. Zhou, and L. Gan, "Robust anomaly sound detection framework for machine condition monitoring," DCASE2022 Challenge, Tech. Rep., July 2022.

[9] J. Jie, "Anomalous sound detection based on self-supervised learning," DCASE2023 Challenge, Tech. Rep., June 2023.

[10] Z. Lv, A. Jiang, B. Han, Y. Liang, Y. Qian, X. Chen, J. Liu, and P. Fan, "Aithu system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[11] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "Eat: Self-supervised pre-training with efficient audio transformer," *arXiv preprint arXiv:2401.03497*, 2024.

[12] 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.

[13] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[14] 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.