XJU SYSTEM FOR UNSUPERVISED ANOMALOUS SOUND DETECTION WITH PRE-TRAINED AUDIO MODELS AND DENOISING

Technical Report

Zhou Yang

XinJiang University
School of Computer Science and Technology
Urumqi, China
yangzhou@stu.xju.edu.cn

Liang He

Tsinghua University
Department of Electronic Engineering
Beijing, China
heliang@mail.tsinghua.edu.cn

ABSTRACT

DCASE 2026 Task 2 focuses on noise-aware unsupervised anomalous sound detection for machine condition monitoring, where systems are required to detect unknown anomalies using only normal training data under noisy and domain-shift conditions. In this technical report, we present the XJU systems based on pre-trained audio model fine-tuning, domain-specific feature learning, denoising-based front-end processing, and ensemble learning. The submitted systems include an OSSCL system, a BEATs-GRL system, an ANC-enhanced system, and an OSSCL-GRL ensemble system. On the development set, the submitted systems outperform the official baselines in terms of the official hmean score, and the ensemble system obtains an hmean score of 65.8%. In addition, we discuss training-time two-channel fusion as a possible future direction for noise-aware anomalous sound detection.

***Index Terms*—** Unsupervised anomalous sound detection, pre-trained audio model, fine-tuning

1. INTRODUCTION

Anomalous sound detection (ASD) is the task of identifying whether the sound emitted from a target machine is normal or anomalous. It is an important technique for machine condition monitoring, since abnormal acoustic patterns may indicate potential faults or changes in machine operating states. In practical industrial environments, anomalous sounds are rare and difficult to collect comprehensively. Therefore, ASD is generally formulated as an unsupervised learning problem, where only normal sounds are available for model training.

DCASE 2026 Task 2 [1][2][3][4]focuses on noise-aware unsupervised anomalous sound detection for machine condition monitoring. Compared with previous tasks, this task emphasizes several practical challenges:

- **Unsupervised learning:** The system must be trained using only normal sounds and detect unknown anomalous patterns that are not observed during training.
- **Domain generalization:** The system is required to handle domain shifts caused by changes in machine operating conditions, recording environments, and environmental noise.
- **Unseen machine types:** The system should be applicable to machine types that are unseen during the development phase, where manual hyperparameter tuning for evaluation machines is not available.
- **Incomplete attribute information:** The system needs to work both when attribute information is available and when it is unavailable.
- **Noise-aware two-channel recording:** The system can exploit synchronized near- and far-microphone recordings to improve robustness against background noise.

These characteristics make DCASE 2026 Task 2 more challenging than a standard unsupervised anomaly detection problem. The system must not only learn discriminative representations of normal machine sounds, but also remain robust to domain shifts, incomplete attribute information, unseen machine types, and noisy multi-channel recording conditions. In particular, the two-channel setting provides additional information from different microphone positions, but it also requires the detection system to effectively use heterogeneous acoustic observations recorded at different distances from the target machine.

In this technical report, we investigate several strategies for DCASE 2026 Task 2. We fine-tune large-scale pre-trained audio models to learn discriminative normal-sound representations, explore a gradient reversal layer based approach to model domain-specific features, introduce an ANC-based denoising front-end to reduce the influence of background noise, and combine complementary systems through ensemble learning. These systems are designed to address the main challenges of noise-aware unsupervised anomalous sound detection under domain-shift and limited target-domain conditions.

2. PRE-TRAINED AUDIO BACKBONES AND FRONT-END PROCESSING

2.1. BEATs

BEATs [5] is a self-supervised audio representation learning framework based on acoustic tokenizers. It learns audio representations by predicting discrete acoustic tokens from masked audio inputs. This training strategy enables the model to capture high-level semantic and contextual information from audio signals. The Transformer encoder in BEATs provides robust frame-level representations that can be transferred to downstream audio tasks. In our submitted systems, BEATs is used as the backbone of the GRL-based system. The pre-trained BEATs encoder is fine-tuned on normal

machine sounds, and the learned representation is further optimized to capture domain-specific acoustic variations under different machine operating conditions and recording domains.

2.2. EAT

Efficient Audio Transformer (EAT) [6]is a self-supervised audio pre-training model designed to learn transferable acoustic representations efficiently. It uses a Transformer-based encoder to model temporal dependencies in audio signals and captures both global and local acoustic information from the input waveform. Compared with models trained from scratch on task-specific data, EAT provides a stronger initialization for anomalous sound detection, especially when the amount of available training data is limited.

In our system, the pre-trained EAT model is used as the backbone of the OSSCL system. The encoder is fine-tuned using normal machine sounds, and the output embedding is optimized to represent different machine conditions in a compact and discriminative feature space.

2.3. ANC Denoising Front-End

In addition to pre-trained audio backbones, we employ an ANC-based denoising front-end to improve the quality of input signals. Background noise, environmental interference, and channel-dependent recording artifacts may affect the extracted acoustic representation and lead to unreliable anomaly scores. Therefore, the ANC front-end is applied to the input waveform before feature extraction.

The denoised waveform is subsequently fed into the anomalous sound detection backend. This design treats ANC[7] as a front-end enhancement module rather than an independent anomaly detector. By reducing irrelevant noise components in the input signal, the ANC-based system aims to improve the robustness of downstream representation learning and anomaly scoring.

3. SUBMITTED SYSTEMS

3.1. System Overview

We submitted four systems for evaluation: an OSSCL system, a BEATs-GRL system, an ANC-enhanced system, and an OSSCL-GRL ensemble system. These systems are designed to investigate different strategies for unsupervised anomalous sound detection, including discriminative representation learning, domain-specific feature learning, denoising-based input enhancement, and ensemble learning.The main configuration and hyperparameters used in our submitted systems are summarized in Table 1. Since the submitted systems include different training and processing strategies, the table reports the primary fine-tuning configuration used in our experiments.

3.2. Common Anomaly Scoring Framework

All submitted systems follow the same anomaly scoring framework. During training, only normal samples are used to optimize the feature extractor. During inference, embeddings extracted from normal training samples are used to construct the reference set for each machine type. The anomaly score of a test sample is computed according to its deviation from the reference embeddings. A larger anomaly score indicates that the test sample is farther from the learned normal distribution and is therefore more likely to be anomalous.

Table 1: Main configuration and hyperparameters of the submitted systems

Parameter	Value
Training Mode	OS-SCL
Loss Function	SubCenterArcFace[8]
Margin Parameter	0.2
Sub-Center Number	114
Total Training Steps	10,000
Batch Size	16
Learning Rate	1×10^{-4}
Optimizer	AdamW[9]
Warm-up Steps	960
Gradient Accumulation	8
Learning Rate Scheduler	inverse sqrt
Data Augmentation	SpecAug[10], 80
Temperature	$t = 0.04$
EMA Momentum	$\alpha_e = 0.9995$
Number of Classes	80
KNN	K=1

3.3. System1

The OSSCL system is based on discriminative fine-tuning of a pre-trained audio model[11]. It uses normal machine sounds to optimize the embedding space with an angular-margin-based classification objective and supervised contrastive learning. The angular-margin objective increases the separation between different machine conditions, while the contrastive objective encourages samples from the same condition to form compact clusters. This system focuses on learning a discriminative representation of normal machine sounds. By constructing a compact and separable embedding space, the OSSCL system provides a strong baseline for unsupervised anomalous sound detection under domain-shift conditions.

3.4. System2

The BEATs-GRL system uses BEATs as the backbone and introduces a gradient reversal layer[12] during training. The pre-trained BEATs encoder provides transferable acoustic representations, while the GRL-based strategy is used to enhance the learning of domain-specific features. This system is designed to model acoustic variations associated with different machine operating conditions and recording domains. Compared with the OSSCL system, the BEATs-GRL system emphasizes domain-related representation learning and provides complementary information for anomaly detection.

3.5. System3

The ANC-enhanced system introduces a denoising front-end before anomalous sound detection. The input waveform is first processed by the ANC module to reduce background noise and recording interference. The enhanced waveform is then fed into the downstream anomalous sound detection system,The other configuration is the same as that of System1.

3.6. System4

The fourth system is an ensemble system constructed from the OSSCL and BEATs-GRL systems. Ensemble learning[13] has been widely used to improve robustness and generalization by combining multiple base systems. In this work, the ensemble system is included as one of the submitted systems.

4. RESULTS

The performance of the submitted systems is evaluated on the development set using the official evaluation metric. The harmonic mean of AUC-s, AUC-t, and pAUC is used as the overall score. The development set results are used for system analysis and parameter selection, while the final ranking of the challenge is determined by the hidden evaluation set.

On the development set, the OSSCL system obtains an hmean score of approximately 63.8%. The BEATs-GRL system obtains an hmean score of approximately 62.8%. The ANC-enhanced system obtains an hmean score of approximately 62.9%. The OSSCL-GRL ensemble system obtains an hmean score of approximately 65.8%. These results provide a reference for comparing the submitted systems under the development-set protocol.

Overall, the development-set results show that the submitted systems exhibit different behaviors under the same evaluation protocol. The OSSCL system represents discriminative fine-tuning of pre-trained audio models, the BEATs-GRL system represents domain-specific feature learning, the ANC-enhanced system investigates denoising-based input enhancement, and the ensemble system is included as one of the submitted systems. Detailed results for AUC-s, AUC-t, pAUC, and machine-level performance will be reported in the final result table2.In addition, the two-channel setting suggests another possible direction beyond the submitted systems. Multi-channel extension strategies, such as inserting cross-channel interaction modules into pre-trained audio encoders, may allow the model to exploit synchronized near- and far-microphone recordings during training. Although this direction is not included in the current submitted systems, it remains a promising extension for noise-aware anomalous sound detection and will be further investigated in future work.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "Beats: audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 5178–5193.

[6] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "Eat: self-supervised pre-training with efficient audio transformer," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024, pp. 3807–3815.

[7] S. M. Kuo and D. R. Morgan, "Active noise control: a tutorial review," *Proceedings of the IEEE*, vol. 87, no. 6, pp. 943–973, 1999.

[8] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou, "Sub-center arcface: Boosting face recognition by large-scale noisy web faces," in *European Conference on Computer Vision*. Springer, 2020, pp. 741–757.

[9] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*.

[10] 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," *Interspeech 2019*, p. 2613, 2019.

[11] S. Huang and L. He, "Xju system for first-shot unsupervised anomalous sound detection," DCASE2025 Challenge, Tech. Rep., June 2025.

[12] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[13] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

Table 2: Performance comparison of systems across machine types

Machine	Metric	baseline-MSE	baseline-MAHALA	System 1	System 2	System 3	System 4
bearingEmu	AUC_s	62.34	65.92	60.98	60.14	61.38	60.82
	AUC_t	59.56	62.28	64.64	62.78	64.14	64.24
	pAUC	59.85	60.42	58.11	60.53	57.00	60.32
fan	AUC_s	61.45	60.00	80.22	58.28	83.96	73.12
	AUC_t	46.94	45.09	43.38	51.96	41.00	45.82
	pAUC	53.33	52.29	49.11	51.05	49.26	51.47
gearboxEmu	AUC_s	68.23	74.48	75.58	63.70	75.92	73.90
	AUC_t	49.78	52.74	66.10	79.10	65.86	78.74
	pAUC	52.94	53.97	58.74	61.00	58.16	59.26
sliderEmu	AUC_s	67.25	66.36	73.12	54.80	71.52	66.04
	AUC_t	45.05	49.18	61.12	56.28	59.84	59.76
	pAUC	50.38	50.36	52.00	49.74	51.79	50.79
ToyCar	AUC_s	75.62	77.28	71.80	75.94	70.00	77.30
	AUC_t	37.87	53.17	78.38	77.78	75.00	82.58
	pAUC	54.03	58.25	61.21	65.16	56.79	66.58
ToyCarEmu	AUC_s	69.62	69.49	66.98	58.84	67.62	68.14
	AUC_t	61.20	66.62	89.50	84.64	90.28	87.14
	pAUC	55.89	53.47	61.11	55.68	59.58	62.21
valveEmu	AUC_s	67.74	56.60	72.34	70.10	72.14	76.84
	AUC_t	68.78	56.50	70.08	88.00	68.18	82.46
	pAUC	55.08	50.20	64.00	69.21	64.53	73.79
Official Score		56.66	57.66	63.86	62.85	62.97	65.80