FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION SYSTEM BASED ON PRE-TRAINED MODEL

Technical Report

Sanghyeok Chung[1], *Sunmook Choi*[2], *Seungeun Lee*[1], *Kihwan Lee*[1], *Il-Youp Kwak*[3], *Seungsang Oh*[1]
[1] Department of Mathematics, Korea University, Seoul, South Korea
[2] Center for Applied Mathematics, Cornell University, Ithaca, NY, USA
[3] Department of Statistics and Data Science, Chung-Ang University, Seoul, South Korea

ABSTRACT

This technical report presents our approach for Task 2 of the DCASE2025 Challenge, First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. To tackle the challenge of detecting anomalous sounds, we utilize a pre-trained model as a feature extractor. We further adapt the model to the task using Low-Rank Adaptation (LoRA), allowing efficient fine-tuning. Anomaly scores are then computed using a k-nearest neighbors algorithm on standardized feature vectors. Experimental results on the development set demonstrate that our proposed system significantly outperforms the official baseline, validating the effectiveness of our approach.

Index Terms— Anomalous sound detection, pre-trained model, domain shift

1. INTRODUCTION

Anomalous sound detection (ASD) refers to the task of determining whether the sound emitted by a target machine is normal or indicates an abnormal condition. The ability to automatically detect mechanical failures through sound analysis plays a crucial role in AI-based factory automation, enabling predictive maintenance and minimizing unplanned downtime. The DCASE2025 Challenge Task 2, *First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring* [1], [2], [3], [4], aims to develop ASD models that can perform well despite environmental or domain shifts, using noisy normal machine sounds and a few clean or noise-only samples for training.

This task involves addressing three major challenges. First, the model must be trained using only normal sound data. Since anomalous events are rare and exhibit diverse patterns, collecting sufficient abnormal samples is inherently difficult. As a result, the model needs to learn how to detect a wide range of potential anomalies using only normal sound during training. Second, the model must detect anomalies robustly under domain shifts. In real industrial environments, changes in machine operating conditions or variations in environmental noise can easily cause domain shifts. To address this, the model should incorporate domain generalization techniques to maintain performance across different domains. Third, the model should generalize to completely new machine types without any additional hyperparameter tuning. Since no evaluation data are available in these situations, conventional tuning strategies cannot be applied. Therefore, the task requires solving a first-shot ASD problem.

In this study, we propose an ASD system specifically designed to tackle the aforementioned three challenges commonly encountered in real-world applications. Our approach leverages a pre-trained audio representation model, BEATs, as the feature extractor. To detect anomalies, we pair this with a k-nearest neighbors (KNN)–based anomaly detection method. We also apply feature standardization during the detection process to make the anomaly scores more consistent, which helps improve the overall performance of the system.

2. METHODOLOGY

Our ASD system is composed of two main components [5], [6]. The first is the feature extractor, which serves as the front-end. This module contains trainable parameters and is used to extract feature vectors from input sound data. The second component is the anomaly detector, which functions as the back-end. This part does not include any trainable parameters. Given the feature vectors produced by the front-end, it calculates the anomaly score for each input.

2.1. Feature extractor

For the feature extraction component, we adopt the pre-trained model BEATs (Bidirectional Encoder representation from Audio Transformers) [7], a self-supervised learning framework designed to learn high-level representations from audio signals. BEATs consists of two main modules: an acoustic tokenizer and an audio SSL (self-supervised learning) model, which are jointly optimized iteratively. The acoustic tokenizer produces semantically meaningful discrete labels, which guide and enhance the representation learning of the audio SSL model.

To adapt BEATs to our task, we fine-tune it using Low-Rank Adaptation (LoRA) [8]. Instead of updating all parameters, LoRA introduces a small number of trainable weights, allowing for efficient fine-tuning. The model is fine-tuned in a classification setting, where each machine type in the dataset is treated as a separate class. For this purpose, we attach a mean pooling layer followed by two dense layers to the BEATs architecture.

We use the BEATs-iter3 version for fine-tuning. Input audio is either zero-padded or truncated to 10 seconds, and then transformed into a log-mel spectrogram using a frame length of 25 ms, a frame shift of 10 ms, and 128 mel bins. We employ the AAM softmax [9] with margin $m = 0.2$ and scale $s = 30$, and optimize the model using the Adam optimizer. LoRA is applied specifically to the query, key, and value projection layers within the self-attention modules of the transformer.

2.2. Anomaly detector

For the anomaly detection component, we employ a KNN approach using cosine distance as the similarity metric. Given a test sample, its feature vector is first extracted by the feature extractor. Then, the distances between this vector and all feature vectors from the training data—belonging to the same machine class—are calculated. The smallest of these distances is used as the anomaly score.

To improve detection performance, we apply feature standardization before computing distances. Specifically, both training and test feature vectors are standardized using the mean and standard deviation computed from the training feature vectors. This normalization step helps ensure consistent distance computation, enhancing the reliability of anomaly scoring.

3. RESULT

The proposed ASD system is evaluated with the AUC and pAUC where $p = 0.1$. Table 1 presents the result of our ASD system. The result shows that our proposed systems outperforms the baseline.

Table 1: Result of submitted systems on the DCASE 2025 task 2 development set

Machine	Metric	Baseline	System1	System2	System3	System4
Valve	AUC(source)	63.53	82.88	76.12	83.6	83.6
	AUC(tartget)	67.18	71.04	72.56	72.52	73.2
	pAUC	57.35	56.57	58.63	65.31	67.78
ToyTrain	AUC(source)	61.76	78.67	76.36	72.36	72.43
	AUC(tartget)	56.46	62.24	66.04	62.15	68.19
	pAUC	50.19	52.31	56.57	53.78	58.78
ToyCar	AUC(source)	71.05	66.79	64.48	60.6	58.75
	AUC(tartget)	53.52	65.32	66.28	71.68	75.04
	pAUC	49.7	51.15	52.68	51.31	51.36
Slide rail	AUC(source)	70.1	82.8	81.16	72.24	77.44
	AUC(tartget)	48.77	58.88	57.68	53.84	52.67
	pAUC	52.32	53.63	53.57	50.73	50.15
Bearing	AUC(source)	66.53	68.24	70.71	70.76	71.6
	AUC(tartget)	53.15	65.36	66.99	62.95	72.52
	pAUC	61.12	54.05	52.42	54.94	57.0
Gearbox	AUC(source)	64.8	74.71	70.43	70.56	67.04
	AUC(tartget)	50.49	74.24	69.08	70.0	55.08
	pAUC	52.49	68.10	54.36	58.94	52.57
Fan	AUC(source)	70.96	50.84	49.47	47.64	46.88
	AUC(tartget)	38.75	66.76	65.12	56.24	55.24
	pAUC	49.46	56.47	55.36	52.31	51.31

4. CONCLUSION

This paper proposes ASD systems for the DCASE 2025 Challenge Task 2. We fine-tune the pre-trained model BEATs by applying LoRA. Then we combine it with KNN detector which performs standardization to the feature vectors. Our proposed systems outperform the baseline.

5. 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*.

[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] 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.

[6] A. Jiang, X. Zheng, Y. Qiu, W. Zhang, B. Chen, P. Fan, W.-Q. Zhang, C. Lu, and J. Liu, "Thuee system for first-shot unsupervised anomalous sound detection," DCASE2024 Challenge, Tech. Rep., June 2024.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[8] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," 2022. [Online]. Available: https://arxiv.org/abs/2212.09058

[9] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, p. 5962–5979, Oct. 2022. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2021.3087709