SUBMISSION FOR THE DCASE2025 TASK2: ROBUST ANOMALY SOUND DETECTION VIA BSS-AUGMENTED PRE-TRAINED MODEL FINE-TUNING

Technical Report

Haifeng Xu[1]*, Yizhou Tan* [2]*, Shengchen Li*[2]

[1] Anhui University of Science and Technology, School of Computer Science and Engineering, Huainan, China, xuuuhf@aust.edu.cn
[2] Xi'an Jiaotong-Liverpool University, School of Advanced Technology, Suzhou, China, Yizhou.Tan22@student.xjtlu.edu.cn, shengchen.li@xjtlu.edu.cn

ABSTRACT

Anomaly Sound Detection (ASD) is crucial for predictive maintenance in industrial settings, yet its performance is often severely constrained by high-intensity, non-stationary background noise. To address this challenge, this paper proposes a robust ASD framework incorporating multi-source data fusion and fine-tuning. Specifically, we fuse machine sounds recorded in factories (containing only normal samples) with easily available clean mechanical sounds or environmental noise data. A pretrained BEATs model serves as the feature extractor. To enhance noise robustness, we innovatively introduce a Blind Source Separation (BSS) decoder as an auxiliary task atop the BEATs encoder. This guides the model in learning feature representations that are resistant to noise interference by minimizing BSS loss. Experiments conducted on the DCASE 2025 Development dataset demonstrate that our method significantly outperforms baseline approaches, achieving AUC values of 79.86% and 71.47% on ToyCar and ToyTrain, respectively. This represents relative improvements of 6.69% and 9.71% over baseline systems, underscoring the efficacy of our proposed framework in acoustic event detection and classification scenarios.

Index Terms— Anomaly detection, Pre-trained model, Blind Source Separation, Fine-tuning

1. INTRODUCTION

Anomaly sound detection (ASD) for industrial machinery is crucial for enabling predictive maintenance and avoiding significant losses caused by equipment failures. Its primary goal is to automatically identify abnormal patterns in machine operating sounds within complex industrial acoustic environments. However, actual factory settings commonly feature high-intensity, non-stationary background noise (e.g., sounds from other equipment, human voices, environmental sounds). This noise significantly degrades the clarity of the target machine sound and masks subtle signs of anomalies, presenting a major obstacle to improving the robustness and real-world deployment capability of ASD systems.

The typical DCASE[1] ASD task setup poses significant challenges: (1) Training data contains only normal sound samples; (2) Significant domain shift exists, meaning test data (target domain) may come from different operating conditions, or environments than the training data (source domain); (3) Validation/test data often consists of entirely new, unseen data types, requiring models to possess strong generalization capabilities. Notably, recent DCASE ASD task rules [citation specific year guidelines] allow participants to use additional audio data collected during training.

Various methods have been proposed for ASD tasks, including AutoEncoders based on reconstruction error[2]. These methods typically train using only normal samples. In recent years, pre-trained audio models[3] (e.g., BEATs[4], CED[5], Unispeech [6]) have demonstrated powerful feature representation capabilities across various audio tasks. Applying these to ASD tasks has become an important direction, with the main strategy being fine-tuning based on the pre-trained models. For example, Jiang et al. proposed a feature consistency-based fine-tuning method at INTERSPEECH 2023 [7]. However, these mainstream fine-tuning approaches usually follow a supervised classification framework. They often require using additional attribute information provided by the dataset (e.g., machine ID, operating load) as classification targets during training[7, 8]. This reliance on detailed attribute labels is a significant limitation in real industrial scenarios, as precisely obtaining and labeling this information is often costly or impractical. Furthermore, the robustness of existing methods under complex noise interference still needs improvement.

To overcome the limitations of current pre-trained model-based ASD methods regarding noise robustness and reduce reliance on hard-to-obtain attribute labels, this paper proposes a novel fine-tuning framework. Making full use of the additional data permitted by DCASE rules, we combine target factory-recorded normal machine sounds with easily collected clean reference machine sounds or environmental background noise. We artificially mix these to generate noisy samples. Innovatively, we introduce Blind Source Separation (BSS) as an auxiliary task to guide the fine-tuning process of a pre-trained model (BEATs). BSS decoder aims to reconstruct clean target machine sound features from the input mixed audio features. By jointly optimizing the loss functions of the BSS auxiliary task, the model is forced to learn deep representations that effectively resist noise interference and focus on the essential acoustic characteristics of the target machine. Crucially, learning the BSS task does not depend on any specific machine attribute labels, enhancing the method's generality.

The remainder of this paper is organized as follows: Section 2 details the proposed method, including data fusion, the BSS auxiliary task framework, and model specifics. Section 3 describes the experimental setup and results analysis. Section 4 concludes the paper and outlines future work.

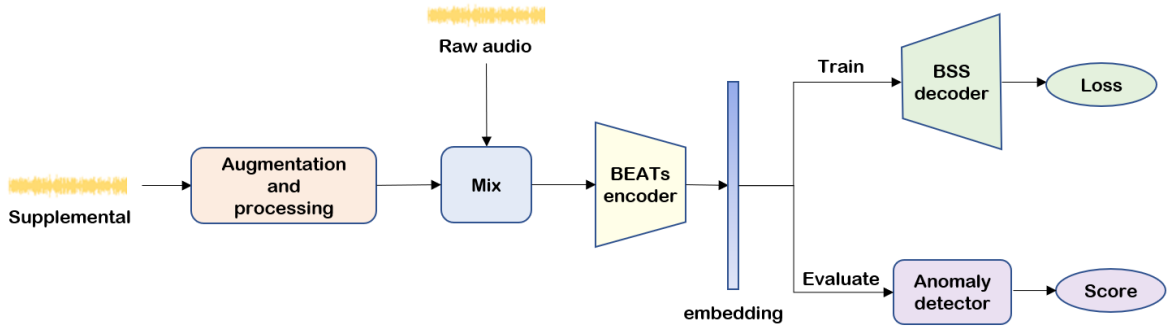

Figure 1: Architecture of our system.

2. PROPOSED METHOD

2.1. Data augmentation

The DCASE 2025 Task 2 dataset[9, 10, 11] provides 1000 clips of normal machine sounds and an additional 100 supplemental clips per machine type. The supplemental clips consist of either clean mechanical sounds (without background noise) or environmental noise recorded when the machine is idle.

To leverage the supplemental data for training our supervised BSS-based model, we artificially corrupt the supplemental clips. Specifically, we generate noisy versions by additively mixing each supplemental clip with randomly sampled segments of Gaussian white noise.

This noise addition process is applied to both types of supplemental clips (clean machine sounds and idle environmental noise). For each of the 100 supplemental clips, we generate 9 distinct noisy variants, resulting in a total of 1000 augmented noisy clips

2.2. BSS-based Supervised Training

Our Anomalous Sound Detection (ASD) model is trained using a self-supervised representation learning approach based on a Blind Source Separation (BSS) proxy task. The core architecture consists of a pre-trained encoder followed by a BSS decoder. During the training phase, solely on normal machine audio data, the model learns robust feature representations (embeddings) by optimizing the BSS decoder to perform a specific task by reconstructing a clean component of the machine sound. The embeddings produced by the encoder serve as the input to this BSS decoder.

To leverage the generalization capabilities learned from large-scale audio datasets, we employ the pre-trained BEATs model as our encoder. This encoder is then fine-tuned on the normal machine audio data alongside the BSS decoder to adapt to the specific acoustic characteristics and optimize its performance for the BSS proxy task.

During the testing phase, the fine-tuned encoder extracts embeddings from input audio segments. A k-Nearest Neighbors (KNN)[12] model is pre-trained on the embeddings computed from the entire training set of normal audio. For a test sample, its embedding is fed into this KNN model. The anomaly score is then derived as the minimum Euclidean distance between the test embedding and its nearest neighbors within the normal training embeddings.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

We trained the model on the development dataset and the additional training dataset. The development dataset contains recordings from 7 machine types that are different from those in the evaluation dataset, while the additional training dataset includes recordings from 8 machine types that are the same as those in the evaluation dataset. All training audio has a sampling rate of 16 kHz. Instead of clipping audio with different lengths, we adjusted the input and output of the network to fit their sizes.

The BSS decoder's loss function is computed conditionally: For supplemental data containing background noise, only the development dataset contributes to the loss.When supplemental data includes machine-generated audio, both development and supplemental data are utilized in loss computation.

To preserve temporal fidelity, the Transformer-based BEATs encoder and BSS decoder jointly process variable length audio sequences without truncation or padding. Model optimization employs the AdamW algorithm with a fixed learning rate of 0.00001, a batch size of 16, and a maximum training duration of 150 epochs. A two stage fine-tuning approach is implemented: (1) *Initial Representation Learning (Epochs 1–50)*: Both encoder and decoder parameters are updated to establish task-specific feature extraction capabilities. (2) *Encoder Refinement (Epochs 51–150)*: The BSS decoder is frozen while encoder parameters are further optimized to enhance anomaly detection performance through domain-invariant representation learning.

During inference, the fully fine-tuned BEATs encoder extracts embeddings from the in-domain training set, which exclusively contains normal operational sounds of target machine types. These embeddings constitute the reference database

3.2. Results

Within the development set, four machine types (ToyTrain, fan, gearbox, and slider) exhibit supplemental data containing only background noise. Consequently, for these machine types, only the development dataset contributes to the loss computation. Conversely, for the remaining three types (ToyCar, bearing, and valve), both development and supplemental data are incorporated into the loss calculation. Table 1 shows the results we achieved on the development set through the BSS-based Supervised Training.

Table 1: :DCASE 2025 Task2 experimental results on development dataset (%).The value in therow "Total Score" represents the harmonic mean of the AUC and pAUC scores over all the machine types,sections,and domains.

	Metric	Baseline (MSE)	Baseline (MAHALA)	Ours
ToyCar	AUC(source)	71.05	73.17	**79.86**
	AUC(target)	53.52	50.91	**57.6**
	pAUC	49.7	49.05	**53.05**
ToyTrain	AUC(source)	61.76	50.87	**71.47**
	AUC(target)	56.46	46.15	**67.24**
	pAUC	50.19	48.32	**55.1**
bearing	AUC(source)	**66.53**	63.63	64.18
	AUC(target)	53.15	**59.03**	52.26
	pAUC	61.12	**61.86**	50.78
fan	AUC(source)	70.96	**77.99**	72.48
	AUC(target)	**38.75**	38.56	28.82
	pAUC	49.46	50.82	**54.21**
gearbox	AUC(source)	64.8	**73.26**	65.3
	AUC(target)	50.49	51.61	**53.16**
	pAUC	52.49	**55.07**	51.78
slider	AUC (source)	70.1	**73.79**	70.47
	AUC (target)	48.77	50.27	**57.44**
	pAUC	52.32	**53.16**	51.1
valve	AUC (source)	63.52	56.22	**68.84**
	AUC (target)	67.18	61	**82.78**
	pAUC	57.35	52.23	**66.99**
Total Score		56.26	55.32	**57.81**

4. CONCLUSIONS

This paper proposes an industrial anomalous sound detection framework that fuses multi-source data with a Blind Source Separation (BSS) auxiliary task. It utilizes the pre-trained BEATs model to extract features and enhances noise resistance through BSS loss optimization. Experiments on the DCASE 2025 development set demonstrate excellent performance, with the AUC values for ToyCar and ToyTrain reaching 79.86% and 71.47% respectively, showing improvements of 6.69% and 9.71% over the baseline. The average overall score increases by 7.12%, validating the effectiveness of the proposed method.

5. REFERENCES

[1] http://dcase.community/challenge2025/.

[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] B. Han, Z. Lv, A. Jiang, W. Huang, Z. Chen, Y. Deng, J. Ding, C. Lu, W.-Q. Zhang, P. Fan, J. Liu, and Y. Qian, "Exploring large scale pre-trained models for robust machine anomalous sound detection," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1326–1330.

[4] 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

[5] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "Ced: Consistent ensemble distillation for audio tagging," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 291–295.

[6] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, and X. Yu, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6152–6156.

[7] V. Zavrtanik, M. Marolt, M. Kristan, and D. Skočaj, "Anomalous sound detection by feature-level anomaly simulation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1466–1470.

[8] X. Zheng, A. Jiang, B. Han, Y. Qian, P. Fan, J. Liu, and W.-Q. Zhang, "Improving anomalous sound detection via low-rank adaptation fine-tuning of pre-trained audio models," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 969–974.

[9] 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.

[10] 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.

[11] 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.

[12] F. Angiulli, S. Basta, S. Lodi, and C. Sartori, "Distributed strategies for mining outliers in large data sets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1520–1532, 2013.