FUSION SYSTEM FOR FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION FOR MACHINE CONDITION MONITORING

Technical Report

Zhe Cao[1], *Jichao Zhang*[1,2], *Xiao-Lei Zhang*[1,2†], *Chi Zhang*[2], *Xuelong Li*[2]

[1] Northwestern Polytechnical University, Xi'an, China
[2] China Telecom, Shanghai, China

ABSTRACT

The DCASE 2025 Challenge Task 2 focuses on first-shot unsupervised anomalous sound detection, where the main challenges include domain shift, generalization issues, and the absence of attribute information. To address these problems, we leverage the newly introduced past-year DECASE Challenge Task 2 datasets and the Audioset for model pre-training to extract audio features. In this work, we employ LoRA fine-tuning, dual-branch feature exchange, and multi-layer feature fusion methods. In addition, data augmentation is utilized to mitigate domain shift, and multiple models are fused to further enhance performance. As a result, an hmean of 67.27% is achieved on the development dataset.

***Index Terms*—** Anomaly detection, pre-trained model, fine-tune, fusion system

1. INTRODUCTION

Anomalous sound detection in industrial equipment is crucial for ensuring the safe operation of machinery. DCASE 2025 Task 2 [1, 2, 3, 4], First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring, focuses on identifying anomalous sounds emitted by given machine types. The proposed methods are required to distinguish between normal and abnormal machine sounds under the interference of complex background noise. The main challenges include:

1. The model is trained only on normal sounds and must detect unseen anomalous events during inference. This reflects real-world conditions where anomalies are rare and diverse.

2. The system must handle changes in machine states or environmental noise without relying on target domain data. Domain generalization techniques are required to maintain performance across varying conditions.

3. The model should work on entirely new machine types without hyperparameter tuning. It must also perform reliably whether or not additional attribute information is available.

Last year's work [5, 6, 7, 8, 9] demonstrated excellent performance in anomalous sound detection by effectively leveraging pre-trained models and domain adaptation techniques. Building upon this strong foundation, we further extended the approach by incorporating additional strategies.

This paper introduces the fusion system, which was developed for Task 2 of the DCASE 2025 Challenge, focusing on unsupervised anomalous sound detection under domain shifts and limited fault data. To address the scarcity of anomalous samples and enhance generalization, we leverage powerful pre-trained models trained solely on normal sound data. The system adopts a dual-branch architecture that integrates features from two distinct models, enabling it to capture complementary acoustic representations. Additionally, we propose a multi-layer feature fusion strategy to combine information from various depths of each model, further improving detection accuracy. Finally, we employ a linear score-level fusion approach to construct four ensemble systems from individual models. The proposed method achieves a final harmonic mean score (hmean) of 67.27%, demonstrating its effectiveness in detecting unknown anomalies in various domains.

Section 2 provides a detailed introduction to the proposed individual models; Section 3 presents our experimental results; Section 4 outlines the submitted systems.

2. MODEL ZOO

2.1. BEATs

BEATs adopts an iterative self-supervised audio pre-training framework based on a bidirectional encoder for learning audio representations. Within this framework, an acoustic tokenizer and an audio SSL model are optimized alternately. The tokenizer discretizes continuous audio features into discrete labels, and the model is trained to predict these labels using a discrete classification loss, which has been shown to outperform conventional reconstruction objectives. In this task, the BEATs-iter3+ variant is employed, which is pre-trained on the full AudioSet [10] dataset and contains 93.88M parameters.

The pre-trained BEATs model is then fine-tuned for machine attribute classification, where each machine type–attribute combination is treated as a distinct class. Additionally, the pure noise and pure machine sound of each machine type are also treated as separate classes.

All audio clips are padded or truncated to a fixed length of 10 seconds and converted into log-mel spectrograms with a frame length of 25 ms, frame shift of 10 ms, and 128 mel bins. The frame-level features generated by BEATs are aggregated via average pooling to obtain utterance-level embeddings, which are then projected to logits using a single dense layer.

Two fine-tuning models are explored: full parameter fine-tuning and LoRA-based fine-tuning. Both use the ArcFace [11] loss and the AdamW [12] optimizer. The former, denoted as FBeats, employs an initial learning rate of 1e-4 and is trained for 40 epochs. The latter, referred to LoRA-based fine-tuning as LBeats, adopts an initial learning rate of 2e-4, is trained for 40 epochs, and applies LoRA modules to the q, v, and out projections of each transformer

† Corresponding author: Xiao-Lei Zhang

Table 1: Performances of single models on the development set

Base	Model	Total	Trainable	bearing	fan	gearbox	slider	ToyCar	ToyTrain	valve	hmean
BEATs	FBEATs	93.88M	93.88M	63.78	62.10	67.11	59.83	58.73	63.96	72.75	63.75
	LBEATs	93.88M	3.58M	68.31	61.41	65.00	60.91	60.64	66.60	74.15	64.99
EAT	FEAT	321.17M	321.17M	68.50	59.79	68.69	66.92	61.12	64.62	78.92	66.46
	LEAT	321.17M	62.91M	62.72	59.03	68.48	67.86	61.28	64.40	71.90	64.83
Dual-Branch system		415.05M	66.49M	65.51	61.53	66.16	64.00	62.05	64.50	75.86	65.39
Fusion system		415.05M	66.49M	65.00	61.99	71.43	67.05	58.86	66.18	73.21	65.92

Table 2: Combination coefficients of four submitted systems

System	FBeats	LBeats	FEAT	LEAT	Dual-Branch	Fusion system
System1	0.1	0.0	0.5	0.0	0.0	0.4
System2	0.0	0.3	0.1	0.3	0.2	0.1
System3	0.1	0.2	0.1	0.2	0.2	0.2
System4	0.1	0.2	0.1	0.2	0.3	0.1

Table 3: Performance of the Four Submitted Systems on the Development Set

Metric	System1	System2	System3	System4
AUC_s	72.57	72.06	72.11	72.15
AUC_t	72.53	71.27	71.57	71.20
pAUC	58.72	57.76	57.89	57.33
hmean	67.27	66.34	66.50	66.16

block with a r of 64. The dense output layer remains fully trainable in both configurations.

To address the limited availability of target-domain data, SMOTE-based [13] oversampling is applied to the target-domain embeddings to match the number of source-domain samples, with the k_neighbors set to 2. Anomaly detection is finally performed using a 1-nearest neighbor classifier based on cosine distance, where the anomaly score of each test sample is defined as the cosine distance to its nearest neighbor in the oversampled target embedding set.

2.2. EAT

EAT is a self-supervised model for audio representation learning. Compared to other pretrained models, it introduces both global utterance-level and local frame-level representations, enhancing the model's ability to understand audio content. Since this year's DCASE task2 allows the use of the full development and external datasets from previous competitions. To enhance the EAT model's capability in extracting machine acoustic voiceprint and improving its generalization performance, the Development and Additional datasets from the DCASE Challenge (2020–2024) Task 2 [4, 3] are combined with the AudioSet dataset to train the EAT pre-trained model. The hyperparameters are consistent with those of the original EAT-larger model, and the model contains 321.17M parameters.

Similar to the fine-tuning strategy used for the BEATs model, both full parameter fine-tuning(FEAT) and LoRA-based fine-tuning(LEAT) versions are also trained for the EAT model. All audio clips are first truncated or padded to 10 seconds, then converted into log-Mel spectrograms with a frame length of 25 ms, frame shift of 10 ms, and 128 Mel-frequency bins. For full fine-tuning , used the AdamW optimizer with an initial learning rate of 1e-5 and train for 40 epochs. During the LoRA fine-tuning stage, the adaptation modules are applied to the q, v, and out projections, with the rank hyperparameter r set to 64, and the dense layers are set as trainable. Finally, k-NN is used for anomaly detection.

2.3. Dual-Branch system

The dual-branch model integrates embeddings from two different models [14], which enhances anomaly detection performance. The CNN models in both branches are replaced with the pretrained BEATs and EAT models, respectively. During the embedding integration, we also introduce feature augmentation using feature exchange (featex). The features from both branches are first exchange and then fused. During the inference stage, the fused features are fed into a knn classifier to compute anomaly scores.

For each sample, we apply a 0.5 probability to decide whether to keep the original features and labels or to use the augmented ones produced through mixup with the flipped samples. The dual-branch model is fine-tuned using LoRA, with the same configuration as used in the individual LoRA fine-tuning of BEATs and EAT. The initial learning rate is set to 2e-4, the model is trained for 40 epochs, and the loss function used is SCAdaCos[15].

2.4. Fusion system

Integrating information from multiple layers in SSL models has proven to be effective. Different layers of a model capture varying levels of semantic information, with lower layers focusing on local patterns and higher layers encoding more abstract representations. By fusing features from multiple layers, the model can leverage complementary information across semantic levels, leading to improved overall performance. Inspired by [16], we concatenate the features from all Transformer layers during the detection phase and use kNN for anomaly detection. For both BEATs and EAT, whether

fully fine-tuned or LoRA fine-tuned, the concatenated features from all layers are used for detection.

For the dual-branch model, feature concatenation is not applied; instead, integrated features from the two branches are directly used. Additionally, we concatenate the final output features from the LoRA fine-tuned BEATs and EAT models, forming our final Fusion system.

3. SUBMITTED SYSTEMS

All of our submitted systems adopt an ensemble structure. First, the anomaly scores from each model are normalized, and then grid search is performed on the development set to find the optimal combination weights. Table 2 shows the combination weights for the four submitted systems, with System 1 achieving the best weights.

Since the EAT model and the dual-branch model performed better during training on the development set, they were assigned higher weights. To balance the ensemble, we manually increased the weight of the BEATs model, resulting in the other system variants.

4. EXPERIMENT RESULTS

The evaluation of the competition is primarily based on the performance metrics of Receiver Operating Characteristic (ROC) Curve(AUC) and partial AUC(pAUC). As shown in Table 1, for each machine type in the development set, we calculated the harmonic mean(hmean) of source-domain AUC, target-domain AUC, and pAUC for each individual model. The best result was achieved by the FEAT model, with hmean of 66.46%.

As shown in Table 3, for each system, we calculated the source-domain AUC, target-domain AUC, pAUC, and their harmonic mean across all machine types in the development set. The best performance was achieved by System 1, with a harmonic mean of 67.27%.

5. CONCLUSION

This paper presents the Fusion system developed for the DCASE 2025 Task 2 challenge. Powerful pretrained models are trained to address the issue of limited mechanical fault data and to enhance the model's generalization capability. A dual-branch architecture is employed to fuse features extracted from two different models, allowing the system to capture more detailed information from the audio. Furthermore, a multi-layer feature fusion strategy is introduced to combine representations from different layers within each model, improving anomaly detection performance. Finally, individual models are integrated into four ensemble systems using a linear combination of anomaly scores. The proposed system achieves a final H-mean score of 67.27%.

6. 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] 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.

[4] 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.

[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] R. Zhao, K. Ren, and L. Zou, "Enhanced unsupervised anomalous sound detection using conditional autoencoder for machine condition monitoring," DCASE2024 Challenge, Tech. Rep., June 2024.

[8] J. Yang, "Adaptive framework for first-shot unsupervised anomalous sound detection in industrial machine monitoring," DCASE2024 Challenge, Tech. Rep., June 2024.

[9] L. Wang, "Two-step anomaly detection: Integrating attribute classification and generative modeling with attribute inference for diverse machine types," DCASE2024 Challenge, Tech. Rep., June 2024.

[10] 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.

[11] 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.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[14] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 276–280.

[15] K. Wilkinghoff and F. Fritz, "On using pre-trained embeddings for detecting anomalous sounds with limited training data," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 186–190.

[16] J. Shi, D. Berrebbi, W. Chen, H.-L. Chung, E.-P. Hu, W. P. Huang, X. Chang, S.-W. Li, A. Mohamed, H.-y. Lee, *et al.*, "Ml-superb: Multilingual speech universal performance benchmark," *arXiv preprint arXiv:2305.10615*, 2023.