XJU SYSTEM FOR FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION

Technical Report

Shun Huang

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

Liang He

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

ABSTRACT

Previous studies have shown that using large-scale audio pre-training models for anomaly sound detection under domain shift scenarios has demonstrated significant promise. In this year's competition, compared to last year, supplementary sets have been added. Due to our lack of understanding in denoising, this dataset was not utilized throughout the training process. In this technical report, we continue to fine-tune large pre-training models, employing sub-center arcface for training, primarily using the BEATs and EAT models. We trained only on the current development set and additional supplementary sets, achieving a score of 64.46% on the development set.

***Index Terms*—** Anomaly detection, fine-tune, sub center arcface

1. INTRODUCTION

Anomalous sound detection plays a critical role in industrial automation by identifying abnormal acoustic signals [1, 2, 3, 4], thereby ensuring the continuous and stable operation of machinery. This year's competition introduces an additional dataset that provides either noise-free operational audio or pure noise samples for each machine, enriching the training and evaluation conditions. The challenge also continues to focus on anomalous sound detection under scenarios with limited attribute information and cross-domain generalization.

Under the condition of limited target domain data across different domains, oversampling techniques such as SMOTE[5] are typically applied during the detection phase to enhance model robustness. For machine categories with limited attribute information, the sub-center ArcFace[6] loss is employed during training. Compared to the conventional ArcFace[7] loss, this approach adaptively pulls different classes toward multiple centers, enabling more robust anomaly detection when partial machine attribute information is unavailable.

2. LARGE PRE-TRAINING MODEL

2.1. BEATs

The BEATs[8] is a self-supervised audio representation learning framework designed to capture rich semantic and contextual information from raw audio signals. Unlike traditional CNN-based approaches, BEATs employs a Transformer-based architecture with bidirectional context aggregation, enabling it to model long-range temporal dependencies in audio sequences. The framework iteratively trains two components: an acoustic tokenizer that generates discrete semantic tokens from audio inputs, and a Transformer encoder that reconstructs these tokens through masked audio modeling. This dual-optimization strategy enhances the model's ability to learn discriminative features for diverse audio tasks. For anomalous sound detection, BEATs leverages its pre-trained weights to extract robust frame-level embeddings, which are further refined through domain adaptation techniques to address cross-machine variability.

2.2. EAT

The EAT[9] introduces a hybrid learning objective that combines global utterance-level and local frame-level audio representations. By integrating contrastive learning at both temporal granularities, EAT achieves superior generalization across varying acoustic conditions. A key innovation lies in its bootstrap-style training paradigm, where the model alternates between generating pseudo-labels and refining its own predictions without requiring human-annotated data. This self-evolving mechanism ensures adaptability to unseen environments. In the context of anomaly detection, EAT's pre-trained encoder serves as a feature extractor that captures subtle deviations in machine operation sounds. Its architecture prioritizes computational efficiency while maintaining sensitivity to rare abnormal patterns, making it particularly suitable for industrial settings with limited labeled data.

3. SUBMITTED SYSTEMS

3.1. Training Configuration

The training parameters for all systems are summarized in Table 1. Key hyperparameters include optimizer settings, learning rate schedules, and adaptation strategies.

We submitted four single-model systems for evaluation, each based on a standalone pre-trained audio model without ensemble strategies. The models are trained exclusively on the training set of DCASE 2025 Task 2.

3.2. System Overview

The submitted systems are categorized into four configurations:

BEATs-FFT 1:The system adopts a dual-branch architecture that combines self-supervised audio representation learning with

Table 1: Training configuration and hyperparameters

Parameter	Value
Training Mode	OS-SCL
Loss Function	SubCenterArcFace[6]
Margin Parameter	0.2
Sub-Center Number	114
Total Training Steps	10,000
Batch Size	16
Learning Rate	1×10^{-4}
Optimizer	AdamW[10]
Warm-up Steps	960
Gradient Accumulation	8
Learning Rate Scheduler	inverse_sqrt
Data Augmentation	SpecAug[11], 80
KNN Settings	R: 0.2, N: 3
Temperature	$t = 0.04$
EMA Momentum	$\alpha_e = 0.9995$
Number of Classes	172

signal-processing-based feature extraction. The first branch is based on the BEATs-iter3 model, a Transformer-based architecture pretrained on large-scale unlabeled audio data. This branch processes the raw waveform and outputs a 128-dimensional embedding that captures high-level semantic and contextual information from the input audio. The second branch performs a full-segment Fast Fourier Transform (FFT) on the audio signal, converting it into the frequency domain. A lightweight convolutional network[12] is then applied to extract low-level spectral features, resulting in an additional 128-dimensional feature vector.These two representations are concatenated into a single 256-dimensional feature vector, which is subsequently used for downstream classification.

EAT-FFT 2: The system adopts a dual-branch architecture that combines the strengths of both self-supervised learning and signal-processing-based feature extraction. The first branch utilizes the EAT-base model, a Transformer-based audio representation learner pretrained on large-scale unlabeled audio data. This branch processes the raw waveform and outputs a 128-dimensional embedding that captures high-level semantic and contextual information. The second branch performs a full-segment Fast Fourier Transform (FFT) on the input audio, converting it into the frequency domain. A lightweight convolutional network is then applied to extract embed, resulting in an additional 128-dimensional feature vector.These two representations are concatenated.

EAT-BEATs 3:The system is a multi-model fusion approach that combines two powerful self-supervised audio representations EAT and BEATs by directly concatenating their output embeddings. Specifically, each input audio is independently processed by both the EAT-base and BEATs-iter3 models, which generate 128-dimensional embeddings capturing complementary acoustic patterns from their respective pre-training objectives. These embeddings are then concatenated into a single 256-dimensional feature vector, which serves as the final representation for downstream classification.

EAT-BEATs-FFT 4:The system extends the **EAT-BEATs** architecture by introducing an additional signal-processing branch based on Fast Fourier Transform (FFT). While **EAT-BEATs** fuses only the embeddings from the EAT and BEATs self-supervised models (256-dimensional total), this system further incorporates a handcrafted spectral feature branch.Specifically, in addition to the 128-dimensional embeddings from both EAT-base and BEATs-iter3, a lightweight CNN processes the FFT-transformed audio spectrogram to extract low-level frequency patterns, producing another 128-dimensional feature vector. Importantly, the features extracted by the FFT branch are explicitly normalized during training. This normalization step helps stabilize the learning process when fusing heterogeneous feature representations.All three representations are concatenated into a final 384-dimensional embedding for downstream classification.

4. REFERENCES

[1] 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.

[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] 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.

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

[6] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou, "Sub-center arcface: Boosting face recognition by large-scale noisy web faces," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 741–757.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 4690–4699.

[8] 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*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202.

Table 2: Performance Comparison of Systems Across Machine Types

Machine	Metric	System 1	System 2	System 3	System 4
bearing	AUC_s	79.48	70.01	78.62	77.02
	AUC_t	72.23	65.94	65.17	67.06
	pAUC	64.94	52.57	62.73	61.57
	hmean	71.73	61.89	68.18	67.97
fan	AUC_s	71.25	68.01	69.88	70.64
	AUC_t	47.14	46.94	44.88	46.86
	pAUC	50.57	50.63	51.15	52.21
	hmean	54.52	53.80	53.43	54.89
gearbox	AUC_s	76.30	77.66	75.60	73.78
	AUC_t	67.74	72.30	67.21	63.58
	pAUC	58.31	60.42	57.68	59.84
	hmean	66.64	69.35	66.02	65.22
slider	AUC_s	75.58	80.40	76.50	75.68
	AUC_t	55.52	54.06	51.66	52.60
	pAUC	54.21	51.21	52.26	52.94
	hmean	60.37	59.44	58.18	58.69
ToyCar	AUC_s	75.58	60.52	66.88	61.89
	AUC_t	55.52	68.74	68.61	71.58
	pAUC	54.21	50.21	51.84	54.05
	hmean	60.37	58.83	61.45	61.69
ToyTrain	AUC_s	76.88	75.70	81.09	80.72
	AUC_t	67.28	69.34	71.82	70.11
	pAUC	56.15	58.15	62.21	56.36
	hmean	65.67	66.92	70.87	67.58
valve	AUC_s	76.97	69.20	75.30	77.84
	AUC_t	75.41	77.30	77.80	81.16
	pAUC	74.73	60.21	62.47	70.21
	hmean	75.70	68.18	71.19	76.12
Average	AUC_s	73.79	71.07	74.55	73.45
	AUC_t	63.44	63.23	61.83	62.75
	pAUC	58.04	54.44	56.75	57.62
	hmean	64.46	62.17	63.55	63.96

PMLR, 23–29 Jul 2023, pp. 5178–5193. [Online]. Available: https://proceedings.mlr.press/v202/chen23ag.html

[9] 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*, ser. IJCAI '24, 2024. [Online]. Available: https://doi.org/10.24963/ijcai.2024/421

[10] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

[11] 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," in *Interspeech 2019*, 2019, pp. 2613–2617.

[12] K. Wilkinghoff, "Design choices for learning embeddings from auxiliary tasks for domain generalization in anomalous sound detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

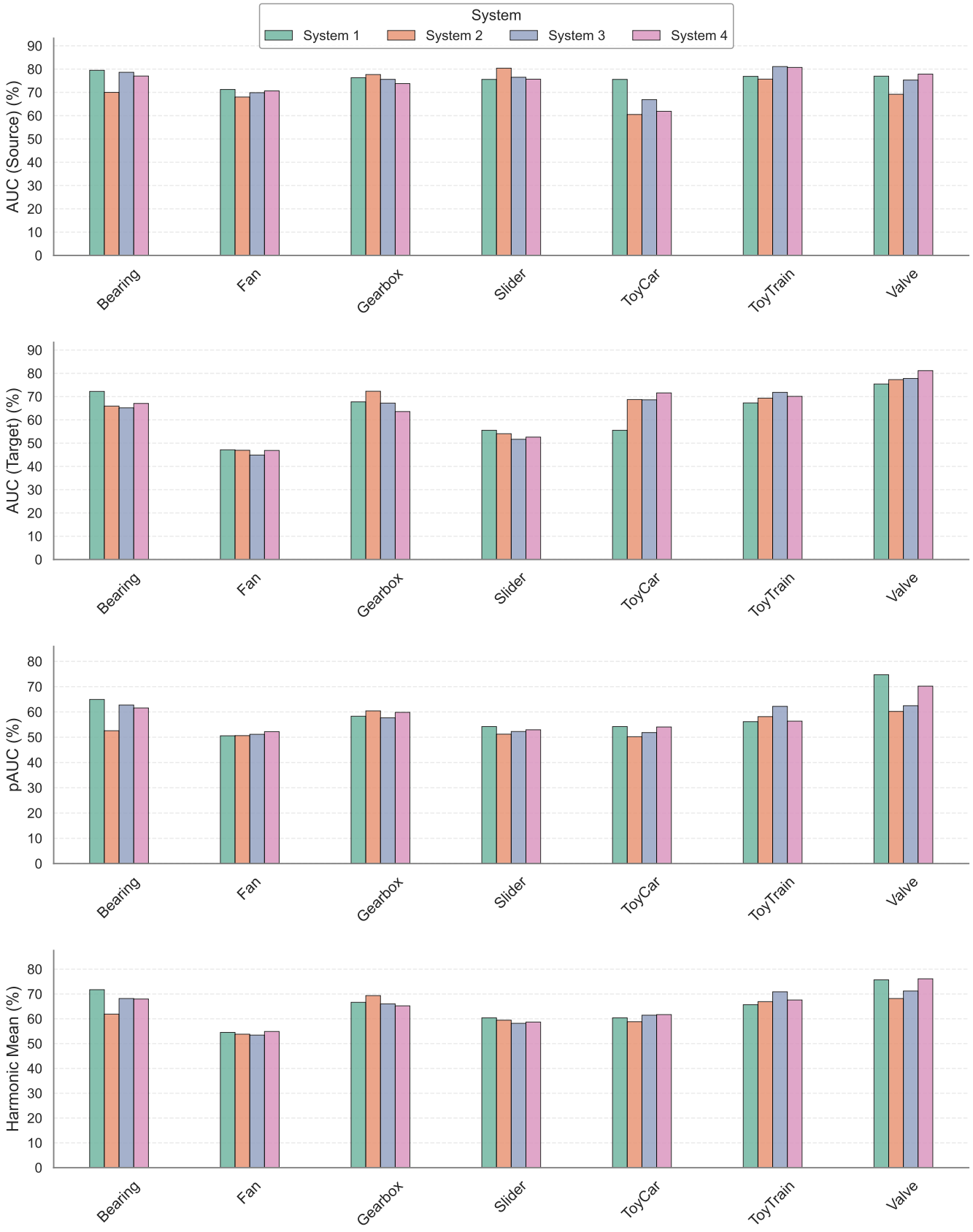

Figure 1: Performance comparison of four systems across different machine types for AUC (Source), AUC (Target), pAUC, and Harmonic Mean metrics.