DUAL-MODE FRAMEWORK FOR FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION IN MACHINE CONDITION MONITORING

Technical Report

Yihao Liu,

Zhejiang ChaoXiLi Technology, Hangzhou, China

865450008@qq.com

ABSTRACT

This technical report presents our solution for the DCASE 2024 Challenge Task 2, which targets First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. This year's task requires the development of a system that functions effectively both with and without attribute information, reflecting real-world scenarios where such data may not always be available. To address this challenge, we propose a dual-mode anomaly detection framework that adapts seamlessly to the availability of attribute information. Our approach leverages advanced pre-training techniques, sophisticated embedding extraction, and refined inlier modeling. In scenarios where attribute information is available, it enhances detection performance; when it is not, the system employs a robust, self-contained strategy to maintain high performance. Our dual-mode system achieves a harmonic mean of 61.598% across all machine types and domains for both the AUC and pAUC ($p = 0.1$) on the development set, demonstrating significant improvement and ensuring versatility under varying data conditions.

Index Terms— Anomalous Sound Detection, Machine Condition Monitoring, Unsupervised Learning, Dual-Mode Framework, Pre-training Techniques, Embedding Extraction, Inlier Modeling, Versatile Detection Systems

1. INTRODUCTION

The Detection and Classification of Acoustic Scenes and Events (DCASE) 2024 challenge includes the task "First-shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring." This task is designed to enhance machine condition monitoring by identifying anomalous sounds that indicate potential faults or unusual behaviors in machinery. Unlike traditional approaches that require extensive labeled datasets of both normal and anomalous sounds, this task employs an unsupervised learning methodology. The model is trained exclusively on normal sound data, making it capable of detecting deviations that signify anomalies [1].

This task addresses several critical challenges in the realm of acoustic anomaly detection:

1. Lack of Labeled Anomalous Data: In real-world industrial settings, acquiring labeled anomalous sound data is often impractical due to the rarity and unpredictability of such events. This task mitigates this issue by using only normal sound data for training [2].

2. First-shot Learning: The task emphasizes the capability to learn from a minimal amount of data. This is particularly important for early-stage fault detection where large datasets are not yet available.

3. Generalization Across Domains: The model needs to generalize well across different domains and machine types, ensuring robustness in diverse operational environments.

4. Building on the foundation of DCASE2023's Task 2, the 2024 task intensifies the complexity by removing additional attribute information of some machines [3].

In this technical report, we will detail our approach to overcoming the challenges of the DCASE2024 "First Shot" task. We plan to leverage pre-trained models, which are designed to extract salient features from audio data. we utilized a meticulously processed AudioSet dataset, with human voice elements removed, to pretrain our models. This strategy enables our models to learn generic sound representations from various acoustic scenes, including those related to machine operations. Then we finetune these pretrained models using Dcase 2024 task2 data. Subsequent to feature extraction using the finetuned models, we introduced anomaly detection algorithms to output anomaly scores.

2. PRE-TRAINED MODELS

In this section we will give a brief description of our pre-trained models using in this task.

1. Audio Spectrogram Transformer (AST): Building upon the success of Vision Transformers (ViT), which are pre-trained on the ImageNet dataset, AST adapts this powerful architecture to cater to the distinct challenges posed by variable-length audio inputs. Unlike ViT, which is designed for fixed-dimension image data, AST is engineered to handle the ever-changing nature of soundscapes. AST employs a scene-centric approach, allowing it to analyze and understand the entire auditory scene rather than just focusing on individual sounds. Furthermore, AST integrates techniques like permutation invariance and sequence masking, which further enhance its ability to distinguish between different sound sources and handle their interplay within complex acoustic environments [4].

2. Self-Supervised Audio Spectrogram Transformer (SSAST): SSAST model utilizes the same architecture as the AST, but with two key differences: While AST employs a [CLS] token to represent the entire audio clip, SSAST uses the mean of all patch embeddings for this purpose which allows for a more aggregated representation of the audio scene; To prevent model leveraging the knowledge of overlapped edges when predicting masked patches, SSAST does not use overlapping patches [5].

3. Bidirectional Encoder representation from Audio Transformers (BEATs): BEATs model revolutionizes audio self-supervised learning (SSL) by moving beyond the conventional mean squared error (MSE) approach. Beats introduces an iterative pre-training framework that jointly learns an acoustic tokenizer and an audio SSL model. In this framework, the acoustic tokenizer generates discrete labels for unlabeled audio data, which are then used to optimize the audio SSL model. Once converged, this model becomes a teacher that guides further training of the tokenizer. This process is repeated until convergence, leading to significantly improved audio representations. By leveraging discrete labels and an iterative teaching strategy, beats taps into the semantic-rich nature of audio more effectively than traditional mse-based methods [6].

Table 1: Anomaly detection results for different machine

	Baseline MSEAE			Baseline MHLAE			Ensemble System		
	AUC(source)	AUC(target)	pAUC	AUC(source)	AUC(target)	pAUC	AUC(source)	AUC(target)	pAUC
ToyCar	66.98%	33.75%	48.77%	63.01%	37.35%	51.04%	61.33%	55.43%	49.43%
ToyTrain	76.63%	46.92%	47.95%	61.99%	39.99%	48.21%	81.55%	56.43%	54.26%
Bearing	62.01%	61.40%	57.58%	54.43%	51.58%	58.82%	68.41%	71.47%	57.44%
Fan	67.71%	55.24%	57.53%	79.37%	42.70%	53.44%	73.53%	41.38%	55.74%
Gearbox	70.40%	69.34%	55.65%	81.82%	74.35%	55.74%	67.24%	66.83%	52.54%
Slider	66.51%	56.01%	51.77%	75.35%	68.11%	49.05%	76.98%	63.66%	55.24%
Valve	51.07%	46.25%	52.42%	55.69%	53.61%	51.26%	87.43%	72.08%	65.69%
All (hmean)	65.00%	50.28%	52.84%	65.77%	49.51%	52.28%	72.85%	59.09%	55.39%

3. METHODOLOGY

3.1. Pre-trained models and classification

Our anomaly detection system consists of two main components. The first is the use of pre-trained feature extractors. The second part involves the statistical anomaly detection algorithms. We observed that machine-generated sounds are more stable, periodic, and have sparser information compared to human speech. Therefore, pre-trained models like wav2vec2.0, Hubert, WavLM, which are trained on speech data, may not be well-suited for extracting features from machine audio. Consequently, we removed the human speech data from the AudioSet and trained models such as AST, SSAST, BEATs. The tasks for these models were sound event classification, with inputs primarily consisting of patches from spectral features which perhaps be better suited for modeling machine characteristics. After completing the pre-training of these models, we finetuned them on the DCASE 2024 Challenge Task 2 dataset with a task of attribute classification to help extract features that are more in line with the characteristics of the training machines.

3.2. Anomaly detection

For the backend anomaly detector, we employed several common algorithms including k-nearest neighbors (KNN), local outlier factor (LOF), Gaussian Mixture Models (GMM) [7] [8] [9], cosine distance, and Mahalanobis distance. We utilized pre-trained models to extract 128-dimensional embeddings from all data, which encompass machine information. Subsequently, we trained our back-end anomaly detection on these embeddings for each machine's training set. Finally, we fed testsets through these models or compared the distance with testsets and trainsets to output anomaly scores. Given that the task emphasizes a "first shot" approach, after observing the performance of the development dataset's test sets, we ultimately selected the KNN model as our system's solution.

3.3. Ensemble

For the competition task emphasizing a "first shot" approach, to achieve more robust results, we integrated the outputs from three pre-trained models. Specifically, we performed z-score normalization for each machine's KNN scores and used weighted sum for the scores of AST, SSAST, and BEATs subsystems to obtain the final system score.

4. RESULTS

Our proposed system was using the development and additional training datasets from DCASE 2024 Task 2. We compared our results against a baseline system employing MSE or Mahalanobis distance on the development test data for seven machines, as depicted in Table 1 [10].

5. CONCLUSIONS

In this technical report, we introduces the system we submitted. We pre-trained models like AST, SSAST, and BEATs using the AudioSet non-human speech data, fine-tuned them with DCASE 2024 task2 data to extract embeddings, and finally output anomaly scores through a trained KNN model. Comparing our experimental results with baselines revealed that our proposed scheme achieved significant performance improvements on most machines. To handle the "first shot" challenge, we integrated the scores of the three models. The system we submitted consists of results with different weight combinations.

6. REFERENCES

[1] Nishida, Tomoya and Harada, Noboru and Niizumi, Daisuke and Albertini, Davide and Sannino, Roberto and Pradolini, Simone and Augusti, Filippo and Imoto, Keisuke and Dohi, Kota and Purohit, Harsh and Endo, Takashi and Kawaguchi, Yohei. Description and Discussion on DCASE 2024 Challenge Task 2: First-Shot Unsupervised Anoma-

lous Sound Detection for Machine Condition Monitoring. In arXiv e-prints: 2406.07250, 2024.

[2] Harada, Noboru and Niizumi, Daisuke and Takeuchi, Daiki and Ohishi, Yasunori and Yasuda, Masahiro and Saito, Shoichiro. 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), 1–5. Barcelona, Spain, November 2021.

[3] Dohi, Kota and Nishida, Tomoya and Purohit, Harsh and Tanabe, Ryo and Endo, Takashi and Yamamoto, Masaaki and Nikaido, Yuki and Kawaguchi, Yohei. 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] Yuan Gong, Yu-An Chung, James Glass. AST: Audio Spectrogram Transformer. In arXiv e-prints: 2104.01778, 2021.

[5] Yuan Gong, Cheng-I Jeff Lai, Yu-An Chung, James Glass. SSAST: Self-Supervised Audio Spectrogram Transformer. In arXiv e-prints: 2110.09784, 2021.

[6] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Furu Wei. BEATs: Audio Pre-Training with Acoustic Tokenizers. In arXiv e-prints: 2212.09058, 2022.

[7] S. Ramaswamy, R. Rastogi, and K. Shim, “Efficient algorithms for mining outliers from large data sets,” in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000, pp. 427–438.

[8] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, Jörg Sander. LOF: Identifying Density-Based Local Outliers. Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX, 2000.

[9] Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6(6), 721-741.

[10] Harada, Noboru and Niizumi, Daisuke and Ohishi, Yasunori and Takeuchi, Daiki and Yasuda, Masahiro. First-Shot Anomaly Sound Detection for Machine Condition Monitoring: A Domain Generalization Baseline. In 2023 31st European Signal Processing Conference (EUSIPCO), 191-195. 2023.