ANOMALOUS SOUND DETECTION USING PRE-TRAINED MODEL WITH STATISTICAL FEATURE DIFFERENCE REPRESENTATION

Technical Report

Shiheng Zhang[1], *Feiyang Xiao*[1], *Shitong Fan*[1], *Qiaoxi Zhu*[2], *Wenwu Wang*[3], *and Jian Guan*[1*]

[1]Group of Intelligent Signal Processing, College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]University of Technology Sydney, Ultimo, Australia
[3]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK

ABSTRACT

This report presents GISP-HEU's submission for Task 2 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2025 Challenge. The submission utilises pre-trained models for feature extraction to obtain refined audio representations. In addition, a statistical weight is formulated based on the differences in audio features between the test and training samples. This weight is applied during the testing phase to enhance the distinction between normal and anomalous audio. The submission comprises four individual systems. System 1 utilises BEATs alongside the statistical feature difference weighting. System 2 builds on System 1 by incorporating clean and noisy data during training. System 3 employs AnoPatch and uses development data spanning DCASE 2022 to DCASE 2025. Finally, System 4 is an ensemble of the previous three systems.

***Index Terms*—** Anomalous sound detection, pre-trained model, statistical clustering, audio representation

1. INTRODUCTION

Unsupervised anomalous sound detection (ASD) focuses on identifying whether the sound emitted by the target machine is anomalous while only normal sounds are available for model training [2, 3]. This is the main topic of the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge Task 2 [4, 5, 6, 7, 8, 1]. In previous DCASE Challenge Task 2, i.e., DCASE 2021 and DCASE 2022, the machine types in the development set are identical to those in the evaluation set. Thus, these methods can adjust hyper-parameters based on the performance on the development set. However, relying on anomalous data to adjust the hyper-parameters of the model is not feasible in real-world scenarios.

First-shot unsupervised anomalous sound detection is introduced in Task 2 of the DCASE 2023 and 2024 Challenges [9, 4, 5, 10, 1]. In this case, the anomalous sounds for the target machine types are not seen during training. As a result, many approaches that depend on adjusting hyper-parameters based on the performance on the development set are no longer applicable to first-shot ASD. Furthermore, DCASE 2025 Challenge introduces the optional use of clean machine or noise-only data for training, enabling participants to enhance model robustness beyond the constraints of noisy operational recordings.

*Corresponding author.

This technical report introduces our submission systems based on statistical learning. Statistical analysis is conducted on the development set to calculate the difference in distribution between the test sample and the training sample, which is then used to amplify the gap between normal and anomalous samples.

Note that the proposed statistical analysis not only considers the distribution difference between the test samples and the training samples, but also the distribution difference between the embeddings of the test samples and training samples. The proposed statistical analysis is adopted to strengthen the audio feature representation during the detection process to distinguish normal and anomalous sounds.

2. PROPOSED SYSTEMS

2.1. System-1: Distribution Difference Weighted System

We use GenRep[11] as our backbone, which adopts the pre-trained model BEATs [12] to extract features and uses the shallow features of each layer for anomaly detection. Inspired by our previous work [2], we assume that the difference in energy distribution between the test set and the training set can reflect the difference between normal samples and abnormal samples. Therefore, a statistical strategy is introduced to enhance feature representation,obtaining the statistical feature difference representation to improve detection performance, and forming the System-1.

2.2. System-2: Distribution Difference Weighted System with Additional Noise-Only and Clean Data

As DCASE 2025 Task 2 allows the use of additional clean machine data and noise-only data as supplementary training resources, we extend System-1 to build System-2 by leveraging these resources to expand the training data used for calculating distribution differences, thereby enhancing feature representation.

2.3. System-3: Data Augmented AnoPatch-Based System

Our third system replicates AnoPatch [13] and extends it by incorporating multi-year training data. Building on AnoPatch, which leverages a ViT backbone pre-trained on AudioSet and fine-tunes it specifically for machine audio, our system emphasizes patch-level modeling to accommodate the inherent sparsity and structure of machine-generated sounds. We further expand the training data by aggregating all development sets from DCASE 2022 to 2025,

Table 1: Performance comparison in terms of AUC-s, AUC-t and pAUC on the development dataset of DCASE 2025 Challenge Task 2.

Methods	ToyCar			ToyTrain			Bearing			Fan			Gearbox			Slider			Valve			Total		
	AUC-s	AUC-t	pAUC	AUC-s	AUC-t	pAUC	AUC-s	AUC-t	pAUC	AUC-s	AUC-t	pAUC	AUC-s	AUC-t	pAUC	AUC-s	AUC-t	pAUC	AUC-s	AUC-t	pAUC	AUC-s	AUC-t	pAUC
AE-MSE [1]	71.05	53.52	49.70	61.76	56.46	50.19	66.53	53.15	61.12	70.96	38.75	49.46	64.80	50.49	52.49	70.10	48.77	52.32	63.53	67.18	57.35	66.78	51.39	52.94
AE-MAHALA [1]	73.17	50.91	49.05	50.87	46.15	48.32	63.63	59.03	61.86	77.99	38.56	50.82	73.26	51.61	55.07	73.79	50.27	53.61	56.22	61.00	52.53	65.51	50.05	52.72
System-1	69.14	75.24	54.65	77.71	73.16	57.42	60.90	68.52	61.51	54.94	58.68	53.06	70.24	70.24	57.79	78.41	56.76	53.80	85.14	61.56	65.50	69.52	65.61	57.39
System-2	68.82	75.44	55.40	78.20	73.52	56.35	60.65	75.04	62.57	55.06	57.80	50.72	70.08	70.64	58.00	75.84	56.16	53.16	85.80	60.00	63.80	69.25	65.96	56.80
System-3	68.29	87.04	61.56	76.65	60.32	50.24	56.57	62.24	52.21	57.88	66.76	52.21	74.73	70.28	59.70	78.41	57.60	52.47	93.67	77.68	72.46	70.38	67.58	56.42
Ensemble System	69.67	84.44	60.71	78.24	71.68	57.36	60.78	73.36	60.29	55.71	61.76	51.09	72.53	71.68	62.52	76.90	57.68	52.79	93.92	67.32	73.15	70.72	68.79	58.99

aiming to increase data availability and enhance the model's generalization capability.

2.4. System-4: Ensemble System

Finally, to take the advantages of each system, we adopt an ensemble learning strategy [14] to integrate System-1, System-2 and System-3, and build an ensemble system. Due to the difference in machine types between the evaluation and development sets, the system weights selected for each machine type on the development set cannot be used on the evaluation set machines. Therefore, we empirically select the same weight for all machine types in our ensemble system.

3. EXPERIMENTS

3.1. Dataset

We conduct experiments on the dataset of DCASE 2025 Challenge Task 2, which comprises a development dataset and an additional dataset [9, 15, 16]. Note that, the machine types in the development dataset are completely different from those in the additional dataset. Our proposed systems are trained on the training set of the development dataset and tested on the test set of the development dataset for effectiveness validation.

3.2. Experimental Setup

For the proposed systems, the machine sound is used with its original sampling rate of 16 kHz. Log-Mel spectrogram is used with a window size of 1024 samples, and overlapping is 50%, where the Mel-filter is set with 128 banks.

3.3. Evaluation Metric

Following the baseline [9], we evaluate our systems using AUC-s, AUC-t, and pAUC metrics. Here, AUC-s and AUC-t represent the Area Under the Curve (AUC) in the source and target domains, respectively, and pAUC denotes the partial AUC. The total AUC-s, AUC-t, and pAUC are computed as the harmonic mean of all machine types.

3.4. Results

We compare our systems with the baseline systems of the DCASE 2024 Challenge Task 2, that is, AE-MSE and AE-MAHALA [9]. The results are given in Table 1, where we can see that all our systems outperform the baseline systems.

4. CONCLUSION

In this technical report, we present our systems for the DCASE 2025 Challenge Task 2, which utilise statistical difference and pre-trained model strategies. Experimental results demonstrate that all our systems outperform the baseline models in first-shot anomalous sound detection.

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*, 2025.

[2] J. Guan, Y. Liu, Q. Zhu, T. Zheng, J. Han, and W. Wang, "Time-weighted frequency domain audio representation with GMM estimator for anomalous sound detection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023.

[3] J. Guan, F. Xiao, Y. Liu, Q. Zhu, and W. Wang, "Anomalous sound detection using audio representation with machine ID based contrastive learning pretraining," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023.

[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 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE Workshop*, Tokyo, Japan, October 2024, pp. 111–115.

[5] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2023 Challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE Workshop*, 2023, pp. 31–35.

[6] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on DCASE 2022 Challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. DCASE Workshop*, 2022, pp. 26–30.

[7] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 Challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in *Proc. DCASE Workshop*, 2021, pp. 186–190.

[8] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on

DCASE 2020 Challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE Workshop*, 2020, pp. 81–85.

[9] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2023, pp. 191–195.

[10] H. Zhang, Q. Zhu, J. Guan, H. Liu, F. Xiao, J. Tian, X. Mei, X. Liu, and W. Wang, "First-shot unsupervised anomalous sound detection with unknown anomalies estimated by metadata-assisted audio generation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 1271–1275.

[11] P. Saengthong and T. Shinozaki, "Deep generic representations for domain-generalized anomalous sound detection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. IEEE, 2025, pp. 1–5.

[12] 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*, 2023, pp. 5178–5193.

[13] A. Jiang, B. Han, Z. Lv, Y. Deng, W.-Q. Zhang, X. Chen, Y. Qian, J. Liu, and P. Fan, "Anopatch: Towards better consistency in machine anomalous sound detection," in *Proc. Interspeech*, 2024, pp. 107–111.

[14] R. L. Sagi Omer, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, 2018.

[15] 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 *Proc. DCASE Workshop*, 2021, pp. 1–5.

[16] 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 ggeneralization task," in *Proc. DCASE Workshop*, 2022, pp. 31–35.