SELF-SUPERVISED ANOMALOUS SOUND DETECTION WITH STATISTICAL CLUSTERING AND CONTRASTIVE LEARNING

Technical Report

Jiantong Tian[1], *Hejing Zhang*[1], *Shiheng Zhang*[1], *Feiyang Xiao*[1], *Qiaoxi Zhu*[2],
Wenwu Wang[3], *and Jian Guan*[1*]

[1] Group of Intelligent Signal Processing (GISP), 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 describes our submission for DCASE 2024 Challenge Task 2. The task aims for first-shot anomalous sound detection and further restricts the use of attribute information (metadata) accompanied by audio signals. This requires the anomaly detection system to function correctly for machine types with and without attribute information. We introduce a statistical clustering strategy to obtain statistical information on audio signals as pseudo-labels to address this issue. In addition, we propose a contrastive learning strategy to enhance audio feature representation by using statistical information, further improving anomaly detection performance. Experiments demonstrate the effectiveness of our proposed strategies, and the results show that all our systems outperform the baseline methods, enabling the model to adapt to the first-shot scenarios without attribute information. Our best system can achieve 71.4% in the harmonic mean of AUC in the source domain, 63.6% in AUC in the target domain, and 56.6% in pAUC.

Index Terms— Anomalous sound detection, contrastive learning, self-supervised learning, statistical clustering, audio representation

1. INTRODUCTION

Unsupervised anomaly sound detection (ASD) focuses on identifying whether the sound emitted by the target machine is abnormal by solely relying on prior knowledge of normal sounds [1–6]. This is the primary focus of Task 2 in the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge [7–10]. In previous DCASE Challenge Task 2, i.e., DCASE 2021, DCASE 2022 and DCASE 2023, the attribute information carried by audios are often used as self-supervised learning labels [2, 3, 11], to enhance the audio representation and improve the anomaly detection performance.

For DCASE 2024 Challenge Task 2 [12], the organizers point out that when recording machine sounds, attribute information regarding the machine condition or noise type are not always available. Therefore, additional attribute information for certain machine types has been hidden. This has rendered previously effective methods unusable for this year's challenge, especially the attribute-based classification methods [2, 11, 13–15].

To this end, we present solutions with statistical information from normal audios and mapping these statistics into a small number of clusters to replace attribute information as pseudo-labels for self-supervised learning, by assuming the statistics are the intrinsic representations of normal audios that can reflect change of acoustic characteristics in domain transfer scenarios. A contrastive learning strategy is also introduced to exploit the relation between audios and statistical clusters to enhance the audio feature representation. Therefore, we can improve the detection performance of the existing state-of-the-art methods with our proposed self-supervised learning and constrastive learning strategies under the situation where attribute information is not available.

2. PROPOSED SYSTEMS

2.1. Systems with Statistical Clustering and Contrastive Learning

Inspired by our submission system for DCASE 2023 Challenge Task 2 [16, 17], this year, we use contrastive learning to establish an intrinsic relation between statistical information and acoustic characteristics. By clustering the statistics of normal audio, we can obtain statistical information as pseudo-labels for self-supervised learning, which constrains the model for better audio representations. Therefore, we adopt two systems as follows:

1. **System-1: Audio feature representation with statistical information and contrastive learning**: We utilize statistical information from normal audio signals to create pseudo-labels and additional labels for self-supervised learning. By using these clustering statistical features as pseudo-labels, we can achieve self-supervised anomalous sound detection without the need of attribute information. Then, a contrastive learning strategy is introduced to exploit the relation between statistical information and acoustic characteristics, thereby enhancing audio feature representation, and enhancing the anomaly detection performance.
2. **System-2: Audio feature representation with metadata information and contrastive learning**: In this system, we adopt attribute information as self-supervised labels for machine types with attribute information. For the machine type without attributes, we only use "source" and "target" as labels, which we found is also effective. This system also em-

[*] Corresponding author.

Table 1: Performance comparison on the development dataset of DCASE 2024 Challenge Task 2, where AUC-s and AUC-t denote source AUC and target AUC respectively, and **Total** means harmonic mean of AUC-s, AUC-t and pAUC over all the machine types.

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
Baseline																								
AE-MSE [12]	67.0	33.8	48.8	76.6	46.9	48.0	62.0	61.4	57.6	67.7	55.2	57.5	70.4	69.3	55.7	66.5	56.0	51.8	51.1	46.3	52.4	65.0	50.3	52.8
AE-MAHALA [12]	63.0	37.4	51.0	62.0	40.0	48.2	54.4	51.6	58.8	79.4	42.7	53.4	81.8	74.4	55.7	75.4	68.1	49.1	55.7	53.6	51.3	65.8	49.5	52.3
Our Systems																								
System-1	67.0	38.7	49.5	71.4	49.2	48.5	75.4	67.3	57.0	76.8	44.5	57.7	71.0	66.4	53.5	68.0	60.4	59.4	56.0	67.1	51.8	68.7	53.8	53.6
System-2	70.1	42.3	49.2	65.2	37.2	49.3	71.9	69.9	59.3	82.1	43.0	59.3	57.7	58.3	51.5	63.3	57.9	50.6	63.2	61.4	57.1	66.9	50.4	53.4
System-3	56.4	48.0	50.0	80.3	61.7	52.4	57.4	67.4	53.1	69.3	52.0	55.5	70.7	69.5	51.1	92.1	86.4	76.3	84.9	79.7	67.1	70.8	**63.9**	**56.7**
System-4	58.2	46.9	49.6	80.4	60.6	52.3	58.8	68.6	53.8	69.7	51.9	55.8	70.5	70.1	51.4	92.2	86.2	76.0	82.9	79.6	65.6	**71.4**	63.6	56.6

ploy our contrastive learning strategy as System-1 to enhance audio representation.

2.2. Self-Supervised Classification System

We also present a self-supervised classification system for Task 2 of DCASE 2024. Specifically, we employ SSL [13] as our backbone and train the classifier using all available labels. For data with attribute groups, we utilize these attributes as classification labels. In cases where attribute information is unavailable, we use the "source/target" domain information as classification labels.

Furthermore, to leverage the information contained in important frequency components, we incorporate an attention module in the frequency dimension [18]. This module is designed to adaptively enhance the important frequency components.

Considering that the audio label may contain important information such as working condition, operating environment, machine type, etc. We propose to transform the labels of audio into descriptive text, and embed them into audio features as auxiliary information. The above steps ensure that the model can refer to attribute group information when learning the audio features.

Finally, for machine types that do not provide attribute information, we also adopt clustering statistical information as pseudo-labels, and use these labels to train the self-supervised classifier for anomaly detection.

2.3. Ensemble System

We employ the ensemble learning strategy [19] to integrate these three systems as our **System-4**. 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 AND RESULTS

3.1. Experimental Setup

Dataset: We conduct experiments on the dataset of DCASE 2024 Challenge Task 2, which comprises a development dataset and an additional dataset [12, 20, 21]. Note that, the attribute information for 3 machine types in the development dataset and 4 machine types in the additional dataset are not provided. Additionally, the machine types in the development dataset are completely different from those in the additional dataset. Our proposed systems (System-1, System-2, and System-3) are trained on the training set of the development dataset and additional dataset for effectiveness validation, and use the weights for the ensemble system, i.e., System-4.
Setup: For System-1 and System-2, the machine sound is used with the original sampling rate of 16kHz, and the learning rate is 0.0001. For System-3, the machine sound is upsampled to 19.2kHz.

Evaluation Metrics: Following the baseline, we evaluate our system using AUC-s, AUC-t, and total AUC metrics. Here, AUC-s and AUC-t represent the Area Under the Curve (AUC) in source and target domain, respectively. pAUC denotes the partial AUC. The total AUC-s, AUC-t, and pAUC is computed as the harmonic mean of all machine types.

3.2. Results

We compare our systems with the baseline systems of the DCASE 2024 Challenge Task 2, i.e., AE-MSE and AE-MAHALA [12]. The results are given in Table 1, where we can see that all of our systems outperform the baseline systems.

4. CONCLUSION

In this technical report, we presented our submission systems for the DCASE 2024 Challenge Task 2. Our proposed systems include two self-supervised systems with statistical clustering and contrastive learning strategies, a self-supervised classification system, and an ensemble system. Experimental results show that our systems significantly outperform the baseline system.

5. REFERENCES

[1] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 816–820.

[2] F. Xiao, Y. Liu, Y. Wei, J. Guan, Q. Zhu, T. Zheng, and J. Han, "The DCASE 2022 Challenge Task 2 system: Anomalous sound detection with self-supervised attribute classification and GMM-based clustering," DCASE 2022 Challenge, Tech. Rep., July 2022.

[3] Y. Wei, J. Guan, H. Lan, and W. Wang, "Anomalous sound detection system with self-challenge and metric evaluation for DCASE 2022 Challenge Task 2," DCASE 2022 Challenge, Tech. Rep., July 2022.

[4] 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. of 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 191–195.

[5] 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. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[6] 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. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[7] 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. of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2023, pp. 31–35.

[8] 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. of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Nancy, France, November 2022, pp. 26–30.

[9] 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. of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Barcelona, Spain, November 2021, pp. 186–190.

[10] 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. of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, November 2020, pp. 81–85.

[11] H. Lan, Q. Zhu, J. Guan, Y. Wei, and W. Wang, "Hierarchical metadata information constrained self-supervised learning for anomalous sound detection under domain shift," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 7670–7674.

[12] 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," *arXiv e-prints: 2406.07250*, 2024.

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

[14] Y. Zhang, J. Liu, Y. Tian, H. Liu, and M. Li, "A dual-path framework with frequency-and-time excited network for anomalous sound detection," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1266–1270.

[15] Z. Liu, Y. Song, X. Zeng, L. Dai, and I. McLoughlin, "DP-MAE: A dual-path masked autoencoder based self-supervised learning method for anomalous sound detection," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1481–1485.

[16] J. Tian, H. Zhang, Q. Zhu, F. Xiao, H. Liu, X. Mei, Y. Liu, W. Wang, and J. Guan, "First-shot anomalous sound detection with gmm clustering and finetuned attribute classification using audio pretrained model," DCASE 2023 Challenge, Tech. Rep., June 2023.

[17] 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. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1271–1275.

[18] H. Zhang, J. Guan, Q. Zhu, F. Xiao, and Y. Liu, "Anomalous sound detection using self-attention-based frequency pattern analysis of machine sounds," in *Proc. of INTERSPEECH*, 2023, pp. 336–340.

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

[20] 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. of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Nancy, France, November 2022, pp. 31–35.

[21] 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. of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, Barcelona, Spain, November 2021, pp. 1–5.