LIGHTWEIGHT SOLUTION USING STATISTICAL ESTIMATION FOR FIRST-SHOT ANOMALOUS SOUND DETECTION

Technical Report

Shiheng Zhang[1], *Hejing 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 presents our submission for Task 2 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2024 Challenge [1]. We introduce statistical strategies to build our lightweight non-deep learning anomalous sound detection (ASD) systems. We analyse the intrinsic statistical characteristics of machine sounds in the time-frequency domain. Then, different statistical information forms weights assigned to the frequency components or the time bins for frequency-weighted or time-weighted audio feature representation, resulting in frequency-weighted and time-weighted ASD systems, respectively. Additionally, the time-weighted system applies SMOTE for data augmentation to mitigate domain shift, which forms an SMOTE time-frequency-weighted ASD system. Finally, we use these systems to build an ensembled ASD system. Experiments show that all four systems achieve better performance than the baseline systems.

***Index Terms*—** Anomalous sound detection, statistical learning, non-deep-learning model, audio representation, lightweight

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 [2–4]. This is the main topic of the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge Task 2 [1, 5–8]. In previous DCASE Challenge Task 2, i.e., DCASE 2021 and DCASE 2022, the machine types in the development set are the same as those in the evaluation set. Thus, these methods can adjust hyper-parameters based on the performance of the development set.

However, relying on anomalous data to adjust the hyper-parameters of the model is not feasible in real-world scenarios. Consequently, first-shot unsupervised anomalous sound detection is introduced in Task 2 of the DCASE 2023 and 2024 Challenges [1, 2, 5]. 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 of the development set are no longer applicable to first-shot ASD.

This technical report presents the lightweight systems of Group of Intelligent Signal Processing (GISP) based on statistical learning. We conduct a statistical analysis of the development set and find that temporal energy information and frequency component information are of significant important in distinguishing normal and anomalous sounds. Based on this analysis and our previous work [3, 9–14], we construct ASD systems using statistical learning to utilize the differences in energy distribution and frequency components in the time-frequency domain for audio feature representation to distinguish normal and anomalous sounds.

2. PROPOSED SYSTEMS

2.1. Energy-Weighted System

We found that there is a significant difference in energy distribution between normal and anomalous machine sounds. According to the description on the official website[1], the test set contains both normal and anomalous audio samples. Thus, we assume that the energy distribution difference between the test set and the training set can reflect the difference between normal and anomalous samples. In this case, we can highlight the difference between normal and abnormal sounds for anomaly detection by utilizing the energy distribution differences between the training set and the test set, without using the anomalous labels in the test set. Therefore, we introduce a statistical strategy to learn the energy distribution difference of the spectrogram of machine sounds as weights assigned to energy feature points, thereby obtaining the energy-weighted audio feature representation and build our Energy-Weighted System.

2.2. SMOTE Energy-Weighted System

Based on Energy-Weighted System, we employ SMOTE [15] to alleviate the domain migration problem, obtaining another system, i.e., SMOTE Energy-Weighted System. By using SMOTE, it can enhance the minority class in the training data to balance the number of samples between the source domain and the target domain, thereby mitigating the domain shift and improving the anomaly detection performance.

2.3. SMOTE Time-Weighted System

We employ our previous study, i.e., Time-Weighted Frequency Representation with Gaussian Mixture Model (TWFR-GMM) [3], as our third statistical system. It introduces a statistical time-weighted

*Corresponding author.

[1] `https://dcase.community/challenge2024/`

Table 1: Performance comparison in terms of AUC-s, AUC-t and pAUC on the development dataset of DCASE 2024 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
Baseline																								
AE-MSE [1]	67.00	33.80	48.80	76.60	46.90	48.00	62.00	61.40	57.60	67.70	55.20	57.50	70.40	69.30	55.70	66.50	56.00	51.80	51.10	46.30	52.40	65.00	50.30	52.80
AE-MAHALA [1]	63.00	37.40	51.00	62.00	40.00	48.20	54.40	51.60	58.80	79.40	42.70	53.40	81.80	74.40	55.70	75.40	68.10	49.10	55.70	53.60	51.30	65.80	49.50	52.30
Proposed Methods																								
Energy-Weighted System (System-1)	55.52	34.82	49.37	75.92	46.82	50.05	53.18	59.28	58.00	75.26	42.36	51.74	82.62	80.24	61.05	88.00	83.46	78.84	76.50	70.00	56.63	70.13	54.14	56.67
SMOTE Energy-Weighted System (System-2)	55.90	35.20	49.11	74.64	49.30	50.37	53.86	64.42	58.95	79.90	31.96	52.74	83.18	80.14	60.79	89.84	85.76	80.26	73.64	71.08	53.74	**70.63**	52.35	56.61
SMOTE TWFR-GMM-Generated (System-3)	62.76	37.34	50.26	67.44	41.98	48.74	55.98	65.44	60.16	73.86	44.58	51.26	76.04	71.26	53.95	85.00	78.48	54.89	70.72	70.84	53.58	69.16	**54.18**	53.05
Ensemble (System-4)	55.46	34.86	49.37	75.90	46.92	50.05	53.30	59.60	58.16	75.54	41.80	51.84	82.70	80.24	61.00	88.06	83.58	79.05	76.40	70.08	56.58	70.18	54.10	**56.71**

frequency representation, which has been proven to be effective in Task 2 of DCASE 2023 [9, 10].

As TWFR-GMM needs to choose different pooling vector weights and the number of mixture components of GMM based on the performance of different machine types on the development set, which is not allowed in first-shot scenario. Therefore, in [9, 10], we use the generated anomalous machine sounds by AudioLDM [16] to select the pooling vector weights of the TWFR-GMM for each machine type. Following our previous studies [3, 9, 10], we build our Time-Weighted System, where SMOTE is also adopted to mitigate domain shift problem.

2.4. Ensemble System [17]

Finally, to take the advantages of each system, we adopt an ensemble learning strategy [17] to integrate System-1 and System-2, 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 can not 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 2024 Challenge Task 2, which comprises a development dataset and an additional dataset [1, 18, 19]. 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 16kHz. Log-Mel spectrogram is used with the 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 [1], 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 source and target domain, respectively, and pAUC denotes the partial AUC. The total AUC-s, AUC-t, and pAUC is 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, i.e., AE-MSE and AE-MAHALA [1]. The results are given in Table 1, where we can see that all of our systems outperforms the baseline systems.

4. CONCLUSION

In this technical report, we presented our non-deep learning lightweight systems using statistical strategies for DCASE 2024 Challenge Task 2. Our submission systems include two energy-weighted systems, a time-weighted system, and an ensemble system. Experiments demonstrate the effectiveness of our proposed statistical solutions for ASD, and results show that all our systems outperform the baseline systems.

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 2024 Challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv e-prints: 2406.07250*, 2024.

[2] 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.

[3] 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. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023.

[4] 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. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023.

[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. Detect. Classif. Acoust. Scenes Events (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. Detect. Classif. Acoust. Scenes Events (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. Detect. Classif. Acoust. Scenes Events (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. Detect. Classif. Acoust. Scenes Events (DCASE) Workshop*, 2020, pp. 81–85.

[9] 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. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 1271–1275.

[10] 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," DCASE2023 Challenge, Tech. Rep., 2023.

[11] 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. INTERSPEECH 2023*, 2023, pp. 336–340.

[12] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2022, pp. 816–820.

[13] 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," DCASE2022 Challenge, Tech. Rep., 2022.

[14] 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. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024, pp. 7670–7674.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research (JAIR)*, vol. 16, pp. 321–357, 2002.

[16] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proc. Int. Conf. Mach. Learn. (ICML)*. IEEE, 2023.

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

[18] 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. Detect. Classif. Acoust. Scenes Events (DCASE) Workshop*, 2022, pp. 31–35.

[19] 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. Detect. Classif. Acoust. Scenes Events (DCASE) Workshop*, 2021, pp. 1–5.