FIRST-SHOT ANOMALOUS SOUND DETECTION WITH GMM CLUSTERING AND FINETUNED ATTRIBUTE CLASSIFICATION USING AUDIO PRETRAINED MODEL

Technical Report

Jiantong Tian[1], *Hejing Zhang*[1], *Qiaoxi Zhu*[2], *Feiyang Xiao*[1], *Haohe Liu*[3], *Xinhao Mei*[3], *Youde Liu*[4], *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] Centre for Audio, Acoustic and Vibration (CAAV), University of Technology Sydney, Ultimo, Australia
[3] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK
[4] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

ABSTRACT

This technical report describes our submission for DCASE 2023 challenge task 2. To address the first-shot and domain shift problem in anomalous sound detection (ASD), we designed an ensemble system that consists of a classification method based on pretrained audio neural networks (PANNs) and a clustering method based on the Gaussian Mixture Model (GMM) with a text-to-audio pretrained model AudioLDM. Experiments on the development set show that our system achieved 77.6% in the harmonic mean of area under curve (AUC) in the source domain, 65.4% in AUC in the target domain, and 56.6% in pAUC across all machine types.

***Index Terms*—** Anomalous Sound Detection, GMM Clustering, Audio Pretrained Model, Self-supervised Learning

1. INTRODUCTION

Unsupervised anomaly sound detection (ASD) aims to detect whether the sound emitted by the target machine is abnormal using only prior knowledge of normal sounds [1–3]. This is the main topic of the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge task 2 [4–6]. 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, methods can adjust their hyper-parameters based on the performance of the development set.

In DCASE 2023 challenge task 2 [7, 8], there is no such issue as the machine types in the development and evaluation sets are entirely distinct, thereby circumventing this problem. Consequently, an approach that works well on the development set may not yield good results on the evaluation set, which is known as the first-shot problem for ASD.

To tackle the first-shot problem of anomaly sound detection, we introduce the state-of-the-art text-to-audio method AudioLDM [9] to learn the general representation of the machine sounds and generate proxy anomaly machine sounds for the adjusting of the ASD systems. Then, we adjust the hyper-parameters of the TWFR-GMM method, which we proposed in [1], to improve the detection performance. In addition, we also include the open-source PANNs [10] module and fine-tune the PANNs module to obtain audio feature representation for ASD.

*Corresponding author.

This work was partly supported by the Natural Science Foundation of Heilongjiang Province with Grant No. YQ2020F010, and the GHfund with Grant No. 202302026860.

2. PROPOSED SOLUTION

2.1. Fine-tuned AudioLDM for Machine Sound

We fine-tune the open-source text-to-audio method AudioLDM [9] on the DCASE 2023 challenge task 2 dataset. Specifically, we expand the metadata information of machine sounds into a text caption as the input for sound generation. Then we use the contrastive language-audio pretraining and diffusion processing to fine-tune the AudioLDM model on the DCASE 2023 challenge task 2 dataset on machine sounds. It allows us to generate the proxy anomalous machine sounds for machine types in the evaluation set, which allows us to address the first-shot problem and adjust the hyper-parameters of the ASD system, i.e., the hyper-parameters in the TWFR-GMM method, to new machine types.

2.2. TWFR-GMM Clustering System

TWFR-GMM [1] selected different pooling vector weights and the number of mixture components of GMM based on performance on the development set for different machine types. However, parameter search based on AUC is unavailable due to the first-shot problem. Therefore, we adopt two systems as follows:

1. **System-1**: We abandon adjusting the hyper-parameters of TWFR-GMM, where the number of the cluster center is set as 1, and the pooling operation is set as maximum pooling for all machine types. In fact, this is the Max-GMM system in the ablation study of [1].
2. **System-2**: We use the proxy anomalous machine sounds generated by AudioLDM to select the pooling vector weights of the TWFR-GMM system for each machine type, where the number of the cluster center is the same as System-1.

In addition, we employ SMOTE [11] in both systems to alleviate class imbalance between the source and target domains in the training data of the development set.

Table 1: Performance comparison in terms of AUC and pAUC on the development dataset of DCASE 2023 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	70.1	46.9	52.5	57.9	57.0	48.6	65.9	55.8	50.4	80.2	36.2	59.0	60.3	60.7	53.2	70.3	48.8	56.4	55.4	50.7	51.2	64.8	49.6	52.8
AE-MAHALA	74.5	43.4	49.2	56.0	42.5	48.1	65.2	55.3	51.4	87.1	46.0	59.3	71.9	70.8	54.3	84.0	73.3	54.7	56.3	51.4	51.1	68.8	52.4	52.4
Proposed Methods																								
Max-GMM (System-1)	66.1	46.2	52.6	59.8	56.7	50.6	58.3	57.7	51.0	71.8	51.2	48.5	73.0	74.5	50.3	96.2	78.4	65.1	96.6	84.4	62.8	71.9	61.3	53.8
AudioLDM-TWFR-GMM (System-2)	73.1	47.7	49.4	56.8	47.0	49.1	60.4	69.7	52.2	69.1	64.9	57.6	85.4	81.0	61.5	97.3	87.7	71.6	98.0	86.7	63.2	74.1	65.3	56.8
Fine-tuned-PANNs (System-3)	55.4	65.9	52.5	65.3	49.2	50.4	77.4	56.6	52.4	70.5	44.8	52.5	56.7	60.6	52.1	89.3	69.4	59.4	60.6	51.9	48.5	66.2	55.7	52.4
Ensemble (System-4)	70.9	55.2	50.2	64.0	49.9	48.9	72.5	62.9	54.8	74.3	59.6	56.1	80.1	78.8	60.6	97.3	86.5	71.1	95.2	83.5	60.1	77.6	65.4	56.6

2.3. PANNs based ASD System

We also propose a PANNs based ASD system as our **System-3**, which introduces the PANNs module, i.e., CNN14 module [10] pretrained on AudioSet [12] for the audio feature representation in ASD. In this PANNs based system, the CNN14 module is used as the audio extractor to extract audio features, and a multi-layer perception (MLP) is used after the CNN14 module as the classifier to predict the metadata information (i.e., attributes group) of machine sounds.

Here, we first load the pretrained parameters of the CNN14 module for the model initialization of the audio extractor in the PANNs based ASD system. Then, we fine-tune the PANNs based ASD system by the attributes group classification operation with the cross-entropy loss function.

After the fine-tuning, we use the audio extractor to obtain the audio embeddings of the machine sounds from the same attributes group, and employ the mean pooling to obtain the attributes group center vector of this attributes group.

In the detection inference stage, we use the audio extractor to obtain the audio embedding of the evaluated sound and then calculate the anomaly score for the anomalous sound detection. Here, the anomaly score is the L2 norm value between the audio embedding of the evaluated sound and its corresponding attributes group center vector.

2.4. Ensemble System

We employ the ensemble learning strategy [13] to integrate the methods proposed above into **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

3.1. Dataset

We conduct the experiments on the dataset of DCASE 2023 challenge task 2, which derives from the ToyADMOS2 dataset [14] and the MIMII DG dataset [15]. This dataset includes a development dataset and an additional dataset. Notably, the machine types in the development dataset are entirely different from those in the additional dataset. We train and validate our proposed systems on the development dataset to verify the effectiveness of System-1, System-2 and System-3 and determine the weights for the ensemble system, i.e., System-4. Then, for the submission, we train and fine-tune our systems on the additional dataset.

3.2. Experimental Setup

For System-1 and System-2, the machine sound is loaded at the original sampling rate of 16kHz, and the SMOTE sampling ratio is set to 0.2 between the source and target domains. For System-3, to fit the requirement of the PANNs parameters, the machine sound is upsampled to 32kHz instead of the original sampling rate 16kHz. We set the learning rate to 0.0001 and employed the cosine annealing strategy.

3.3. Evaluation Metric

Following the baseline [7], we employ the AUC-s, AUC-t, and the total AUC metrics for the evaluation. Here, AUC-s denotes AUC in the source domain, AUC-t denotes AUC in the target domain, and pAUC denotes partial AUC. The total AUC is the harmonic mean of AUC across all machine types.

3.4. Results

We compare our systems with the baseline systems of DCASE 2023 challenge task 2, i.e., the AE-MSE and the AE-MAHALA. All of our systems outperform the baseline systems, as shown in Table 1.

4. CONCLUSION

In this technical report, we introduce our submission systems to DCASE 2023 challenge task 2. Our submission systems include a Max-GMM based system, a TWFR-GMM based system, a PANNs based system and an ensemble system. The experiments show that all of our submission systems outperform the baseline systems.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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*, Online, November 2020, pp. 81–85.

[7] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," *arXiv e-prints: 2303.00455*, 2023.

[8] 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," *arXiv e-prints: 2305.07828*, 2023.

[9] 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. of the International Conference on Machine Learning (ICML)*. IEEE, 2023.

[10] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[11] 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.

[12] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[13] R. L. Sagi Omer, "Ensemble learning: A survey," *Wiley interdisciplinary reviews. Data mining and knowledge discovery*, vol. 8, 2018.

[14] 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.

[15] 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.