UNSUPERVISED ABNORMAL SOUND DETECTION BASED ON FEATURE FUSION IN FIRST-SHOT CONDITION

Technical Report

Yao Xiao[1]*, Tao Peng*[1]*, Shi Feng*[1]*, Yanli Wang*[2]*, Hao Ba*[2]*,*
Chenyang Zhu[1]*, Shengchen Li*[3]*, Xi Shao*[1]*,*

[1] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China, {1022010304, 1021010411, 1222013924, chenyangzhu, shaoxi}@njupt.edu.cn
[2] SAMSUNG Electronics(China) R&D Centra, Nanjing, China, {yanli08.wang, hao.ba}@samsung.com
[3] School of Advanced Technology, Xi'an Jiaotong-liverpool University, Suzhou, China, {Shengchen.Li}@xjtlu.edu.cn

ABSTRACT

The DCASE2023 Challenge Task2 is to develop an unsupervised detection system of anomalous sounds for seven types of machines under first shot conditions. In this paper, we use a novel feature fusion way as the system input, using two simple models: one is Autoencoder(AE) and another is GMM. It shows that our feature fusion has significantly improved the results compared with the baseline in general, especially the GMM.

Index Terms— Anomalous sound detection, Feature fusion

1. INTRODUCTION

The DCASE2023 Challenge Task 2 named "First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring"[1][2] focused on the first-shot problem. The first-shot problem can be defined that the set of machine types in the development dataset and evaluation dataset are completely different. The purpose of this task is for the organizers to develop a robust system which can be adapted to any machine.

In the past we have used denoising algorithm based on Deep Xi [3] [4] and model using the classification based model, but none of these methods are applicable in this year's dataset. Considering the special characteristics of first-shot tasks, we uses two simple models to solve the first-shot problem.

2. PROPOSED SYSTEM

2.1. Audio processing

We use spectral coherence to extract audio features and generate 129 $\times$ 696 feature matrix as input. On this basis, the audio features were extracted by mel-spectrogram and wavelet packet energy spectrum, and the three audio features were combined into 257 $\times$ 696 dimension eigenmatrix.

2.2. Feature extraction

2.2.1. Spectral coherence

Spectral coherence estimation is based on short-time Fourier transform (STFT), which evidences periodic energy flows in and across frequency bins for a cyclostationary signal.The Fourier transform of the interactions of the STFT coefficients can returns a quantity which scans the spectral correlation along cyclic frequency axis [5]. Then two-dimensional spectral coherence maps are obtained, which are utilized as one of our feautre inputs.

2.2.2. Wavelet packet energy

Wavelet packet energy feature is based on the audio signal in the time-frequency resolution space features of energy distribution of signal is the essential attribute of division, with clear physical meaning, wavelet packet energy feature has a strong ability to resist noise, can choose the most critical key features of structure in the group, so as to reduce the dimension of feature vector, We generate a 128 $\times$ 1 dimensional feature array based on audio data.

2.2.3. log-Mel

Furthermore, we extract a 128*313 logMel eigenmatrix from audio based on the Baseline system.

2.3. maching learning model

2.3.1. AE based system

According to this year's dataset some audio existed with the machine in a non-operational state, we selected the Autoencoder to get the frame level features, and the frame level features were used to calculate the anomaly score.

2.3.2. Clustering based system

According to this literature, simple clustering algorithms can also achieve good results in the task of anomalous sound detection. Therefore, we uses GMM based model to do this task.

3. EXPERIMENTAL SETUP

3.1. Dataset

The dataset used for our system consists of MIMII DG [6] and ToyADMOS2 [7], which contains normal and abnormal sounds from seven real machines, Fan, Gearbox, Bearing, Slider, ToyCar, ToyTrain, and Valve. Each piece of audio is 10 seconds of single-channel audio, including sounds from machines and related equipment as well as ambient sounds. But in the evaluation dataset, the sets of machine types are completely different from the development dataset. The datasets consist of sounds from seven types of real/toy machines. Each recording is single-channel audio, including a machine's operating sound and environmental noise. The duration of recordings varies from 6 to 18 sec, depending on the machine type. It is worth to be noted that each type of machine consists of one "section".

3.2. Result

The performance of our system is given in Table 1. To evaluate the performance of our method, the anomaly scores are translated into AUC value [8]. AUC is defined as the area enclosed by the coordinate axis under the ROC (Receiver Operating Characteristic) curve. The AUC for each machine type, section, and domain are defined as:

$$\mathrm{AUC}_{m,m,d} = \frac{1}{N_- N_+} \sum_{i=1}^{N_-} \sum_{j=1}^{N_+} \mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \quad (1)$$

$$\mathrm{pAUC}_{m,m,d} = \frac{1}{\lfloor pN_- \rfloor N_+} \sum_{i=1}^{\lfloor pN_- \rfloor} \sum_{j=1}^{N_+} \mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \quad (2)$$

where m represents the index of a machine type, n represents the index of a section, $d = \{\mathrm{source}, \mathrm{target}\}$ represents a domain, $\lfloor \cdot \rfloor$ is the flooring function, and $\mathcal{H}(x)$ returns 1 when x > 0 and 0 otherwise. Here, $\{x_i^-\}_{i=1}^{N_-}$ and $\{x_j^+\}_{j=1}^{N_+}$ are normal and anomalous test clips in the domain d in the section n in the machine type m, respectively. N_- and N_+ are the numbers of normal and anomalous test clips in the domain d in the section n in the machine type m, respectively.

Table 1: Performance of the proposed system

Domain	Bearing	fan	gearbox	slider	valve	ToyCar	ToyTrain
AUC-S	59.8	85.18	84.53	95.64	91.76	70.78	72.82
AUC-T	66.64	53.92	81.21	80.10	79.60	51.1	62.02

4. CONCLUSION

In this technical report, we have presented our system submitted for DCASE2023 challenge Task 2, which uses a novel feature fusion way as the feature input. The results show that all our systems can outperform the baseline systems.

5. REFERENCES

[1] 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 preprint arXiv:2305.07828*, 2023.

[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," *arXiv preprint arXiv:2303.00455*, 2023.

[3] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *1can*, vol. 111, 2019.

[4] Q. Zhang, A. M. Nicolson, M. Wang, K. Paliwal, and C. X. Wang, "Deepmmse: A deep learning approach to mmse-based noise power spectral density estimation," *2can*, vol. PP, no. 99, pp. 1–1, 2020.

[5] J. Antoni, G. Xin, and N. Hamzaoui, "Fast computation of the spectral correlation," *Mechanical Systems and Signal Processing*, vol. 92, pp. 248–277, 2017.

[6] 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 generalization task," in *4can*, Nancy, France, November 2022.

[7] 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 *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 1–5.

[8] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "Auc: a misleading measure of the performance of predictive distribution models," *Global ecology and Biogeography*, vol. 17, no. 2, pp. 145–151, 2008.