UNSUPERVISED ABNORMAL SOUND DETECTION BASED ON SPECTRAL COHERENCE AND FEATURE FUSION IN DOMAIN DISPLACEMENT CONDITION

Technical Report

Tao Peng[1]*, Rui Qiu*[1]*, Junyi Zhu*[1]*, Yao Xiao*[1]*, Su Wang*[2]
Yipeng Zhang[2]*, Chenyang Zhu*[1]*, Shengchen Li*[2]*, Xi Shao*[1]*,*

[1] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China, {1021010411, 1220013123, 1221013732}@njupt.edu.cn
xy13923895999@163.com, chenyangzhu2018@163.com, shaoxi@njupt.edu.cn
[2] School of Advanced Technology, Xi'an Jiaotong-liverpool University, Suzhou, China,
{su.wang19, yipeng.zhang20, Shengchen.Li}@xjtlu.edu.cn

ABSTRACT

The DCASE2022 Challenge Task2 is to develop an unsupervised detection system of anomalous sounds for seven types of machines under domain shifted conditions. In this paper, two systems are proposed: one only uses spectral coherence as feature input and another combines spectral coherence, wavelet and log Mel. It shows that three-feature fusion has significantly improved the results compared with the baseline in general, but sometimes spectral coherence alone can lead to better results. Therefore, we suggest to use both methods in order to get stable results.

Index Terms— DCASE2022 Task2, Unsupervised anomalous sound detection, Domain shifted conditions, Spectral coherence, Feature combining

1. INTRODUCTION

The DCASE2022 Challenge Task 2 named "Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques" [1]focus on solving the problems caused by domain shifts. Domain shifts are differences in acoustic characteristics between the training and test data caused by discrepancies in operational speed, machine load, and environmental noise.In order to solve this problem, we believe that the problem needs to be solved at the source, that is, processing the audio to mitigate the effect of domain shift or adopting better features to characterize the audio.

In DCASE2021 task2, we uesd a denoising network based on Deep Xi [2, 3] as a preprocess for removing noise from the original samples and reducing the effect of conditions changes by treating it as a type of noise.Although this improves the performance of anomaly detection, it also makes the system more complex.Therefore, in this year's work, we aim to find a way to balance model complexity and anomaly detection performance.

The abnormal sound of the machine is mostly caused by the malfunction of the machine. The cyclic spectral analysis algorithm can effectively extract the periodic fault modulation information hidden in the cyclostationary signal through the correlation between the signals, which reflects the unique advantages of the cyclic spectral analysis algorithm compared with the traditional signal processing methods [4]. At the same time, the correlation operation also has the function of reducing noise, which makes the spectral correlation algorithm show a high resistance to noise. Therefore, we believe that the spectral coherence can achieve two effects simultaneously in terms of noise reduction and better representation of machine sound features without using any pre-trained model.

In our work this year, we used spectral coherence as features first together with MobileNetV2 [5] as anomaly detection model, which achieved good results. On this basis, feature fusion is applied combining spectral coherence, log-Mel spectrum and wavelet packet energy spectrum, which leads to even better results on some machines.

2. PROPOSED SYSTEM

An overview of the proposed system which is separated into feature extraction, training and testing phases, as shown in Figure 1 and 2. The procedure of the proposed method is described in detail in the following sections.

2.1. Audio processing

We use spectral coherence to extract audio features and generate 129×396 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 257×396 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 [6]. Then two-dimensional spectral coherence maps are obtained, which are utilized as one of our feautre inputs.

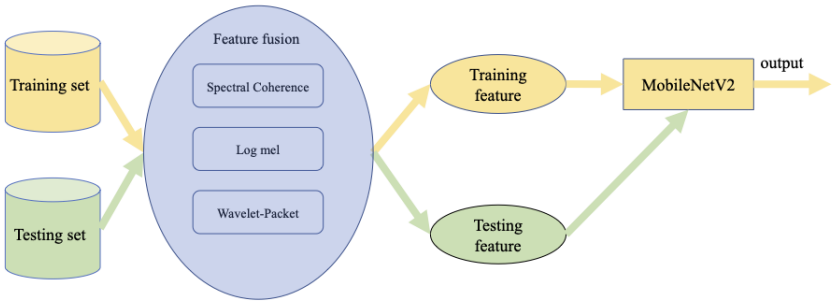

Figure 1: System overview-feature combining.

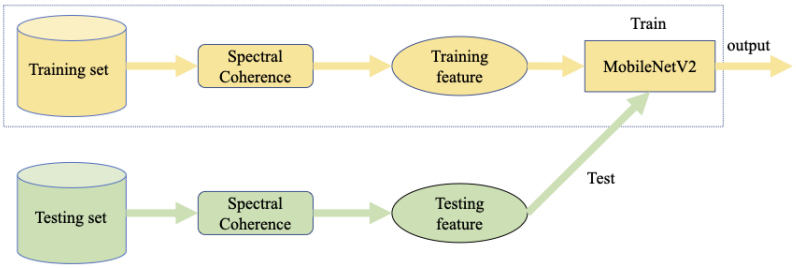

Figure 2: System overview-spectral coherence.

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.

2.2.4. *Feature combining*

In the second system submitted by us, the spectral coherence, wavelet packet energy feature and logMel spectrum feature extracted by us are combined in matrix dimension.

2.3. Classifier

This part, we follow the DCASE2022 MobileNetV2 baseline but to train an overfitting model for each type of machine by leveraging the information of section.

The learning task is to create classification boundary for each section. It identifies from which section the observed signal was generated. In other words, it outputs the softmax value that is the predicted probability for each section. Due to the overfitting of the normal data, in test phase, the output of abnormal data will have a large difference with that of normal one.

The off-the-shelf Keras implementation of MobileNetV2 is used with the width multiplier parameter set to 0.5. The loss function is categorical cross-entropy and the optimization algorithm is adam with 10^{-5} learning rate. The batch size is 32 with 30 epochs, the split percentage of validation is 0.1 after data shuffle.

2.4. Outlier Detection

In this work, the anomaly score is calculated as the averaged negative logit of the predicted probabilities for the correct section, which can be described as:

$$A_\theta(X) = \frac{1}{B}\sum_{b=1}^{B} log\left\{\frac{1-p_\theta(\varphi_{t(b)})}{p_\theta(\varphi_{t(b)})}\right\}, \quad (1)$$

where B is the num of frames, $t(b)$ is the beginning frame index of the b-th image, φ is the acoustic feature, and p_θ is the softmax output by MobileNetV2 for the correct section.

To determine the anomaly detection threshold, assuming A_θ follows the gamma distribution. The parameters of the gamma distribution are estimated from the histogram of A_θ, and the anomaly detection threshold is determined as the 90th percentile of the gamma distribution. If A_θ for each test clip is greater than this threshold, the clip is judged to be abnormal; if it is smaller, it is judged to be normal.

3. EXPERIMENTAL EVALUATION

3.1. Dataset

The data set used for our system consists of MIMII DUE [7] and ToyADMOS2 [8], 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. Each machine has three sections, each of which is a complete set of training and test data. For each section, the data set provides (I) approximately 1000 normal sound fragments in the source domain for training; (ii) Only three normal sounds in the target area are used for training; (iii) About 100 fragments of normal and abnormal sounds in the source domain for testing, and (iv) about 100 fragments of normal and abnormal sounds in the target domain for testing.

3.2. Evaluation metrics

To evaluate the performance of our method, the anomaly scores are translated into AUC value and pAUC value. AUC [9] is defined as the area enclosed by the coordinate axis under the ROC (Receiver Operating Characteristic) curve. pAUC is calculated as the AUC over a low false-positive-rate (FPR) range $[0, p]$. In this task, $p = 0.1$. The AUC and pAUC for each machine type, section, and domain are defined as:

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

$$\mathrm{pAUC}_{m,m,d} = \frac{1}{\lfloor pN_- \rfloor N_+}\sum_{i=1}^{\lfloor pN_- \rfloor}\sum_{j=1}^{N_+}\mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \quad (3)$$

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.

3.3. Experiment Results

Table 1: Detailed results for Fan.

Feature	evaluate	00	01	02	a mean	h mean
log-Mel	AUC(source)	71.07	76.26	67.29	71.54	70.75
	AUC(target)	62.13	35.12	58.02	51.76	48.42
	pAUC	55.40	52.14	65.14	57.56	56.90
SC	AUC(source)	62.66	60.06	75.54	66.08	65.43
	AUC(target)	43.18	63.44	56.82	54.48	53.08
	pAUC	58.47	55.26	56.42	56.71	56.68
Combining	AUC(source)	36.84	86.62	70.90	64.80	56.83
	AUC(target)	30.98	65.08	57.58	51.21	46.14
	pAUC	52.05	69.84	66.42	62.77	61.74

Table 2: Detailed results for Gearbox.

Feature	evaluate	00	01	02	a mean	h mean
log-Mel	AUC(source)	63.54	66.68	80.87	70.37	69.21
	AUC(target)	67.02	66.96	43.15	59.04	56.19
	pAUC	62.12	56.85	50.62	56.53	56.03
SC	AUC(source)	63.34	52.36	67.14	60.94	60.26
	AUC(target)	58.20	50.78	49.32	52.76	52.49
	pAUC	53.21	52.36	58.36	54.64	54.52
Combining	AUC(source)	69.90	62.80	73.74	68.81	68.50
	AUC(target)	69.68	72.18	46.32	62.72	60.24
	pAUC	61.10	62.68	51.57	58.45	58.02

Table 3: Detailed results for ToyCar.

Feature	evaluate	00	01	02	a mean	h mean
log-Mel	AUC(source)	47.40	62.02	74.19	61.21	59.12
	AUC(target)	56.40	56.38	45.64	52.81	51.96
	pAUC	49.96	50.92	56.51	52.46	52.27
SC	AUC(source)	48.02	59.76	57.34	55.04	54.54
	AUC(target)	66.18	47.80	58.80	57.59	56.56
	pAUC	52.57	49.68	50.89	51.05	51.02
Combining	AUC(source)	43.58	57.42	39.04	46.68	45.47
	AUC(target)	71.36	66.20	83.86	73.80	73.09
	pAUC	50.78	55.63	62.57	56.33	55.92

The experimental results are shown in the following table. We propose two kinds of systems: one uses spectral coherence only as feature input, another combines spectral coherence, wavelet packet energy and logMel spectrum to generate eigenmatrix. The baseline system's MobileNetV2 is used for training. The experimental results are compared with the Mel feature used in baseline, and the feature fusion result of slider and valve machine has a great improvement, but bearing and fan have little, so we enhanced the data of bearing and fan. However, the fan's results from spectral-coherence-only system are pretty well. We suspect that feature fusion may weaken the effect of spectral coherence for some machines under some conditions, therefore, we retain the results of spectral coherence for comparison. It's worth noting that none of the methods used in previous years worked well in the target domain of machines like ToyCar, but the feature fusion approach improved the target domain by 20 percent.

Table 4: Detailed results for slide.

Feature	evaluate	00	01	02	a mean	h mean
log-Mel	AUC(source)	87.15	49.66	72.70	69.84	65.15
	AUC(target)	80.77	32.07	32.94	48.59	38.23
	pAUC	71.57	48.21	49.69	56.49	54.67
SC	AUC(source)	95.88	96.06	90.08	94.01	93.92
	AUC(target)	85.26	83.34	61.04	76.54	74.79
	pAUC	66.84	61.52	62.26	63.54	63.45
Combining	AUC(source)	98.34	97.86	90.56	95.58	95.45
	AUC(target)	86.22	84.88	69.50	80.20	79.43
	pAUC	73.47	69.47	67.89	70.28	70.20

Table 5: Detailed results for valve.

Feature	evaluate	00	01	02	a mean	h mean
log-Mel	AUC(source)	75.26	54.78	76.26	68.77	67.09
	AUC(target)	43.60	60.43	78.74	60.92	57.22
	pAUC	55.37	54.69	85.74	65.27	62.42
SC	AUC(source)	82.34	85.90	80.42	82.89	82.82
	AUC(target)	76.80	91.84	70.94	79.86	78.93
	pAUC	64.68	83.52	70.26	72.82	72.00
Combining	AUC(source)	98.04	76.80	89.10	87.98	87.09
	AUC(target)	94.08	67.36	65.00	75.48	73.42
	pAUC	85.31	62.68	59.42	69.14	67.41

4. SUBMISSIONS

In this report, we present two abnormal sound detection systems. Both are based on DCASE2022 MobilenetV2-based Baseline. But the input audio features are modified:one system utilizes spectral coherence, another took three-feature fusion as input.

5. CONCLUSION

In this paper, two abnormal sound detection systems are proposed to perform DCASE 2022 task 2: one uses spectral coherence as feature input and another combines spectral coherence, wavelet and log Mel. By applying MobilenetV2 classifier, our new systems performs much better than the baseline. It shows that three-feature fusion has significantly improved the results compared with the baseline in general, but sometimes spectral coherence alone can lead to better results. Therefore, we suggest to use both methods in order to get stable results.

Table 6: Detailed results for bearing.

Feature	evaluate	00	01	02	a mean	h mean
log-Mel	AUC(source)	67.85	59.67	61.71	63.07	60.58
	AUC(target)	60.17	64.65	60.55	61.79	59.94
	pAUC	54.41	55.09	64.18	57.89	57.14
SC	AUC(source)	61.76	79.68	73.68	71.70	70.89
	AUC(target)	52.62	78.82	66.04	65.82	64.05
	pAUC	51.47	66.73	61.10	59.77	59.08
Combining	AUC(source)	78.46	81.06	71.84	77.12	76.91
	AUC(target)	63.72	65.80	63.04	64.18	64.16
	pAUC	58.89	63.05	54.15	58.70	58.47

Table 7: Detailed results for ToyTrain.

Feature	evaluate	00	01	02	a mean	h mean
log-Mel	AUC(source)	46.02	71.96	63.23	60.40	57.26
	AUC(target)	49.41	45.14	44.34	46.30	45.90
	pAUC	50.25	52.97	51.54	51.59	51.52
SC	AUC(source)	70.96	74.12	66.62	70.56	70.43
	AUC(target)	36.36	44.52	51.56	44.14	43.25
	pAUC	49.57	51.94	54.57	52.03	51.95
Combining	AUC(source)	67.08	71.9	65.52	68.16	68.06
	AUC(target)	42.32	46.90	72.08	53.76	50.99
	pAUC	48.79	50.26	53.73	50.92	50.84

6. REFERENCES

[1] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, 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," *arXiv preprint arXiv:2206.05876*, 2022.

[2] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, 2019.

[3] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "Deepmmse: A deep learning approach to mmse-based noise power spectral density estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1404–1415, 2020.

[4] J. Antoni, "Cyclostationarity by examples," *Mechanical Systems and Signal Processing*, vol. 23, no. 4, pp. 987–1036, 2009.

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

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

[7] 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 arXiv e-prints: 2205.13879*, 2022.

[8] 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.

[9] 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.