COMPARATIVE EXPERIMENTS ON SPECTROGRAM REPRESENTATION FOR ANOMALOUS SOUND DETECTION

Technical Report

*Kazuki Morita**, *Tomohiko Yano**, *Khai Q. Tran**

Intelligent Systems Laboratory, SECOM CO.,LTD.
{morita-ka,tomo-yano,ku-chan}@secom.co.jp

ABSTRACT

In this paper, we propose an anomalous sound detection method for DCASE2022task2. This is the task of anomalous sound detection for machine condition monitoring, and it is required to detect unknown anomalous sound only from normal sound data. Our system is based on a submission system for DCASE2021task2, and we newly evaluated variations in the time-frequency representation used in anomalous sound detection. As a result, the proposed method showed a detection performance of 84.80% for source domain and 82.26% for target domain in Area Under Curve (AUC) and 68.65% in partial AUC (pAUC).

***Index Terms*—** Anomalous Sound Detection, Convolutional Neural Network, Per-Channel Energy Normalization, Harmonic-Percussive Source Separation

1. INTRODUCTION

Anomalous sound detection is the task of identifying whether the sound emitted from a target machine is normal or anomalous. DCASE2022task2[1] includes scenarios for domain shifts caused by maintenance and differences in machine's physical parameters, in environmental conditions and in recording devices.

In DCASE2021task2, we proposed a method which applied self-supervised learning to extract embedding features of a CNN for computing anomaly score using LOF or k-NN[2], and it was found that representative features were more important in anomalous sound detection. Therefore in DCASE2022task2, we compared and evaluated variations of spectrogram representations for acoustic features.

The rest of this report is organized as follows. Chapter 2 describes our anomalous sound detection method. Chapter 3 describes the evaluation experiments. Chapter 4 describes the conclusion. Chapter 5 describes our submission systems.

2. ANOMALOUS SOUND DETECTION METHOD

2.1. Audio Processing

We transform all audio clip into spectrograms. The frame size for STFT is 128 ms, and hop size is 32 ms. We set these parameters experimentally. We apply Per-Channel Energy Normalization(PCEN)[3] and Harmonic-Percussive Source Separation(HPSS)[4] to the spectrograms.

2.1.1. Per-Channel Energy Normalization(PCEN)

PCEN normalizes a spectrogram representation by performing automatic gain control to improve robustness to loudness variation. We apply PCEN to spectrograms, and experimentally set the parameter based on the indoor application described in [3].

2.1.2. Harmonic-Percussive Source Separation(HPSS)

HPSS decomposes a spectrogram into harmonic and percussive components. We use the harmonic and percussive components separately.

2.2. Feature extractor

By using spectrograms and attribute indices in source domain, we train a MobileFaceNet(MFN)[5]. Additionally, we use Additive Angular Margin Loss[6] as a loss function. A spectrogram of 1024 dimensions × 32 frames is used as a processing unit, and the unit is shifted by 16 frames in the audio clip. Model structure is shown in Table 1. As a result, we obtain a 128 dimensions vector per a unit.

Table 1: MobileFaceNet Architecture

Input	Operator	t	c	n	s
$1024\times32\times1$	conv2d 3×3	-	64	1	2
$512\times16\times64$	depthwise conv2d 3×3	-	64	1	1
$512\times16\times64$	bottleneck	2	64	5	2
$256\times8\times64$	bottleneck	4	128	1	2
$128\times4\times128$	bottleneck	2	128	6	2
$64\times2\times128$	bottleneck	4	128	1	2
$32\times1\times128$	bottleneck	2	128	2	1
$32\times1\times128$	conv2d 1×1	-	512	1	1
$32\times1\times512$	linear GDConv16×1	-	512	1	1
$1\times1\times512$	linear conv2d 1×1	-	128	1	1

2.3. Anomaly Detector

We apply k-Nearest Neighbors(k-NN)[7], Local Outlier Factor(LOF)[8] and Gaussian Mixture Model(GMM)[9] in order to calculate anomaly score. We merge embedding vectors in an audio clip using mean or standard deviation

2.4. k-Nearest Neighbors(k-NN)

This method is based on the distance of k-neighboring features. In k-NN, the larger the distance to the selected neighborhood, the more deviated from normal. In this report, we use the mean of cosine distance as the anomaly score, and we set the number of neighbors to 1.

*Equal contribution.

2.5. Local Outlier Factor(LOF)

This method is based on local density, which is the density of k-neighboring feature values. When a feature is anomalous, the difference is large between the local density of the anomaly and the neighboring feature. In this report, we use the outputs of LOF as the anomaly score. We set the number of neighbors to 4.

2.6. Gaussian Mixture Model(GMM)

This method is based on the likelihood of GMM. We estimate parameters of the GMM by using training features. We then calculate the likelihood of the test feature. When a feature is anomalous, the likelihood is small. In this report, we use negative log-likelihood as the anomaly score. We set number of mixture components to 1, and the co-variance type to full.

3. EVALUATION EXPERIMENTS

3.1. Experimental Condition

10-sec length audio (monaural, 16 kHz) was sampled from machinery sound sources. There are seven types of machines (Machine Type); ToyCar, ToyTrain[10], bearing, fan, gearbox,slider and valve[11]. For each Machine Type, there are 3 sections in development dataset and 3 sections in additional dataset. We trained a CNN using 6 sections datasets in a Machine Type, and an anomaly detector using embedding vectors per section. We used librosa[12], scikit-learn[13] and PyTorch Lightning[14] for the implementation. In the experiment, we evaluated the application of PCEN and HPSS to spectrograms.

3.2. Results

The results are shown in Table 2, Table 3, Table 4. Table 2 shows AUC for source domain and Table 3 shows AUC for target domain. Table 4 shows partial AUC for source domain and target domain. Each value is a harmonic mean overall sections.

4. CONCLUSION

In this paper, we used the normal sound of the machine and its attribute index to train a CNN in a self-supervised learning manner. Then, we detected anomalous sound by using feature vectors extracted from the CNN. We evaluated the effectiveness of PCEN and HPSS applied to the input to the CNN per Machine Type. We showed the performance of 84.80% for source domain and 82.26% for target domain in Area Under Curve (AUC) and 68.65% in partial AUC (pAUC).

5. SUBMISSIONS

In this report, we submit four anomalous sound detection systems. Table 5 shows the conditions we used.

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," *In arXiv e-prints: 2206.05876*, 2022.

[2] K. Morita, T. Yano, and K. Tran, "Anomalous sound detection using cnn-based features by self supervised learning," DCASE2021 Challenge, Tech. Rep., July 2021.

[3] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J. P. Bello, "Per-channel energy normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2019.

[4] J. Driedger, M. Müller, and S. Disch, "Extending harmonic-percussive separation of audio signals," in *Proceedings of the 15th International Conference on Music Information Retrieval (ISMIR)*, Taipei, Taiwan, 2014, pp. 611–616.

[5] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices," pp. 428–438, 2018.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[7] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. ACM, 2000, pp. 427–438.

[8] M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. ACM, 2000, pp. 93–104.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[10] 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.

[11] 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.

[12] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[14] W. Falcon et al., "Pytorch lightning," *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, vol. 3, 2019.

Table 2: Harmonic Mean of AUC in the source domain of Development Dataset(%)

spectrogram representation	Detector	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve	all class
baseline(AE)		91.41	76.32	54.45	78.59	68.94	77.95	52.04	68.84
baseline(MNv2)		58.97	58.59	62.88	71.35	69.58	66.03	67.75	64.68
spectrogram	k-NN	90.46	85.03	61.78	82.00	84.42	94.70	99.40	83.67
spectrogram	LOF	88.43	84.23	51.73	75.94	83.30	90.82	99.16	78.85
spectrogram	GMM	83.94	70.93	60.17	76.11	81.22	90.64	95.21	78.11
PCEN	k-NN	90.42	86.20	62.43	82.82	89.52	94.59	98.18	84.66
PCEN	LOF	91.42	84.79	64.84	70.33	82.71	89.58	96.59	81.41
PCEN	GMM	86.62	74.44	63.35	82.99	79.02	92.35	89.76	80.06
HPSS(harmonic)	k-NN	88.13	83.40	62.43	76.96	86.74	85.18	95.91	81.36
HPSS(harmonic)	LOF	83.90	82.12	58.17	71.85	87.96	72.43	91.73	76.68
HPSS(harmonic)	GMM	89.26	71.05	53.89	71.28	80.28	87.90	76.85	73.91
HPSS(percussive)	k-NN	90.29	84.41	55.00	73.08	82.37	93.69	99.41	79.85
HPSS(percussive)	LOF	88.79	85.13	52.26	64.85	77.64	92.10	98.98	76.58
HPSS(percussive)	GMM	84.05	75.43	49.84	62.11	81.64	92.03	92.54	73.50
Our Best		90.46	86.20	62.43	82.99	89.52	94.59	99.16	84.80

Table 3: Harmonic Mean of AUC in the target domain of Development Dataset(%)

spectrogram representation	Detector	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve	all class
baseline(AE)		35.01	23.51	58.66	47.23	62.64	47.70	49.47	42.05
baseline(MNv2)		52.26	46.07	61.81	48.53	56.60	40.72	58.01	51.06
spectrogram	k-NN	88.65	52.13	67.11	74.52	85.73	86.94	88.69	75.09
spectrogram	LOF	68.51	46.71	54.00	58.50	67.51	74.99	96.06	63.54
spectrogram	GMM	86.28	49.43	74.10	69.48	86.83	78.60	73.29	71.73
PCEN	k-NN	87.18	68.38	74.47	73.06	87.01	88.86	88.78	80.25
PCEN	LOF	73.53	62.92	64.68	67.43	77.28	72.69	91.63	71.87
PCEN	GMM	86.98	64.53	67.34	77.27	80.80	77.66	58.65	72.10
HPSS(harmonic)	k-NN	88.71	54.14	76.37	73.63	82.53	80.76	84.10	75.42
HPSS(harmonic)	LOF	65.64	48.23	58.15	62.86	81.74	56.87	77.30	62.59
HPSS(harmonic)	GMM	81.93	52.89	75.11	66.86	79.33	76.77	40.95	63.97
HPSS(percussive)	k-NN	85.74	53.97	70.60	64.14	79.33	83.98	88.72	73.12
HPSS(percussive)	LOF	66.36	49.84	67.51	59.42	63.64	62.82	93.28	64.15
HPSS(percussive)	GMM	83.67	48.71	64.73	63.92	79.37	78.11	61.83	66.57
Our Best		88.65	68.38	76.37	77.27	87.01	88.86	96.06	82.26

Table 4: Harmonic Mean of pAUC in the source domain and the target domain of Development Dataset(%)

spectrogram representation	Detector	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve	all class
baseline(AE)		52.76	50.50	52.03	57.53	58.50	55.78	50.36	53.75
baseline(MNv2)		52.39	51.56	57.35	57.10	56.18	54.77	62.70	55.80
spectrogram	k-NN	71.91	50.36	58.21	63.83	75.64	72.99	90.64	66.95
spectrogram	LOF	64.78	49.69	52.68	55.52	64.03	69.68	88.66	61.51
spectrogram	GMM	59.44	50.49	56.55	57.07	72.35	67.32	66.88	60.62
PCEN	k-NN	71.00	57.24	55.99	65.34	72.98	74.07	85.61	67.57
PCEN	LOF	69.48	58.90	57.96	59.15	67.24	66.97	78.34	64.76
PCEN	GMM	61.35	58.21	58.66	64.80	67.85	70.37	63.13	63.21
HPSS(harmonic)	k-NN	71.73	51.59	59.98	61.64	69.65	65.41	81.56	64.74
HPSS(harmonic)	LOF	60.00	51.75	52.95	53.78	69.75	53.27	67.54	57.67
HPSS(harmonic)	GMM	61.00	51.57	51.57	59.55	66.70	64.31	55.45	58.07
HPSS(percussive)	k-NN	70.14	51.43	55.85	57.81	62.73	67.68	91.62	63.33
HPSS(percussive)	LOF	63.58	51.12	53.33	56.45	55.68	62.78	83.93	59.55
HPSS(percussive)	GMM	58.06	50.89	53.24	54.26	65.07	70.24	64.42	58.72
Our Best		71.91	57.24	59.98	64.80	72.98	74.07	88.66	68.65

Table 5: Submission of our System

System name	Spectrogram representation	Anomaly detector	score merge
Morita_SECOM_task2_1	spectrogram	k-NN	std(valve), mean(otherwise)
Morita_SECOM_task2_2	PCEN	k-NN	std(valve), mean(otherwise)

Morita_SECOM_task2_3	spectrogram(ToyCar),
	HPSS-harmonic(bearing),
	HPSS-percussive(valve)
	PCEN(othewise)
		k-NN	std(valve), mean(otherwise)

Morita_SECOM_task2_4	spectrogram(ToyCar, valve), HPSS-harmonic(bearing), PCEN(othewise)	LOF(valve), GMM(fan), kNN(otherwise)	std(valve), mean(otherwise)