UNSUPERVISED DETECTION OF ANOMALOUS MACHINE SOUND USING VARIOUS SPECTRAL FEATURES AND FOCUSED HYPOTHESIS TEST IN THE REVERBERANT AND NOISY ENVIRONMENT

Technical Report

Jihwan Park and Sooyeon Yoo

Advanced Robot Research Laboratory, LG Electronics, Seoul, South Korea
{jihwan.park, sooyeon.yoo}@lge.com

ABSTRACT

In this technical report, we describe our anomalous sound detection (ASD) systems submitted in DCASE 2020 Task2. To improve the ASD performance in the reverberant and noisy condition, normal machine sound augmentation, focused hypothesis test, and selecting the distinctive spectral features is applied to deep neural network (DNN)-based autoencoder (AE). In the experiments, we found that our approaches outperform baseline methods under the condition that only reverberant and noisy normal sound samples have been provided as training data.

Index Terms— DCASE Challenge 2020 Task2, Anomalous sound detection, Unsupervised learning, Deep neural networks, Autoencoder, Data augmentation

1. INTRODUCTION

Recently, unsupervised learning-based anomaly sound detection (ASD) methods have been actively researched. Even though, a large-scale database is essential for training and fairly evaluating sound detection algorithm, in real-world, anomalous sounds rarely occur and are highly diverse. Therefore, massive patterns of anomalous machine sounds are impossible to deliberately make and/or collect. This means we have to detect unknown anomalous machine sounds that were not reflected in the given training data.

Most ASD systems adopt outlier detection techniques because it is difficult to collect a massive amount of anomalous machine sound data. In [1], deep neural network-based autoencoders (AE) have been adopted to build up ASD systems. Acoustic feature is extracted from encoder part of AE, and then input vector reconstructed at the decoder part of AE. By using reconstruction error of AE, defined as mean square error (MSE) between input and reconstructed vector, statistical hypothesis test result can be computed with pre-defined threshold value [2]. In [3]-[4], AE-based acoustic feature extractor can be optimized to maximize the true positive rate under an arbitrary false positive rate by adopting the Neyman-Pearson Lemma. Furthermore, in [4], the authors proposed an outlier sampling algorithm in latent vector space to artificially generate anomalous machine sounds in order to increase the difference of hypothesis test results between normal and anomalous sounds.

In Detection and Classification of Acoustic Scenes and Events (DCASE) 2020 Challenge [2], unsupervised detection of anomalous sounds for machine condition monitoring has been launched. This task also provides a freely accessible machine sound database [5]-[6], which consists of normal and anomalous operating sounds of six types of toy and real machines. To improve the accuracy of anomalous machine sound detection, we used relevant spectral features such as linear- and log-scaled spectrogram and used data in a focused range of the hypothesis test for evaluating our submitted systems. With the provided development set, we achieved the area under the receiver operating characteristic (ROC) curve (AUC) of 81.79% and the partial-AUC (pAUC) of 68.55%.

2. REVIEW OF THE BASELINE SYSTEM

The AE-based baseline system [2] of DCASE Challenge task2 consists of 9 fully-connected layer where batch normalization and rectified linear unit (ReLU) activation function are applied for all layers except for the output layer. Each machine sound sequence is converted to the time-frequency domain by using short-time Fourier transform (STFT) with frame size of 1024 and half overlap. After that, every 5 consecutive STFT coefficients are fed into a mel-filter bank to obtain 128 dimension log mel spectra feature vector. Input feature vector $\mathbf{x}$ can be reconstructed as follows:

$$\hat{\mathbf{x}}_\tau = D\{E\{\mathbf{x}_\tau|\theta_E\}|\theta_D\} \tag{1}$$

Here, E and D denote the encoder and decoder parts of AE respectively, and θ_E and θ_D correspond to model parameters. After that, by using reconstructed feature vector $\hat{\mathbf{x}}_\tau$, anomaly score $\mathcal{A}$ is defined as the MSE between input $\mathbf{x}_\tau$ and reconstructed vector $\hat{\mathbf{x}}_\tau$ as given by

$$\mathcal{A}(\mathbf{x}, \hat{\mathbf{x}}) = E\{||\mathbf{x} - \hat{\mathbf{x}}||_2^2\} \tag{2}$$

where $E\{\cdot\}$ and $||\cdot||_2$ denote mathematical expectation and L_2 norm, respectively. Finally, anomaly score is classified when the score $\mathcal{A}$ exceeds threshold value.

3. SUBMITTED SYSTEMS

In this section, we describe our submitted ASD systems using several techniques including normal machine sound augmentation, various spectral features, and focused hypothesis test in order to improve ASD performance in the reverberant and noisy conditions. The simplified block diagram of our proposed system is depicted in Fig. 1. Each method is described in the following subsections.

3.1. Normal machine sound augmentation

The AE-based ASD system can be improved by increasing the difference of anomalous scores between normal and anomalous ma-

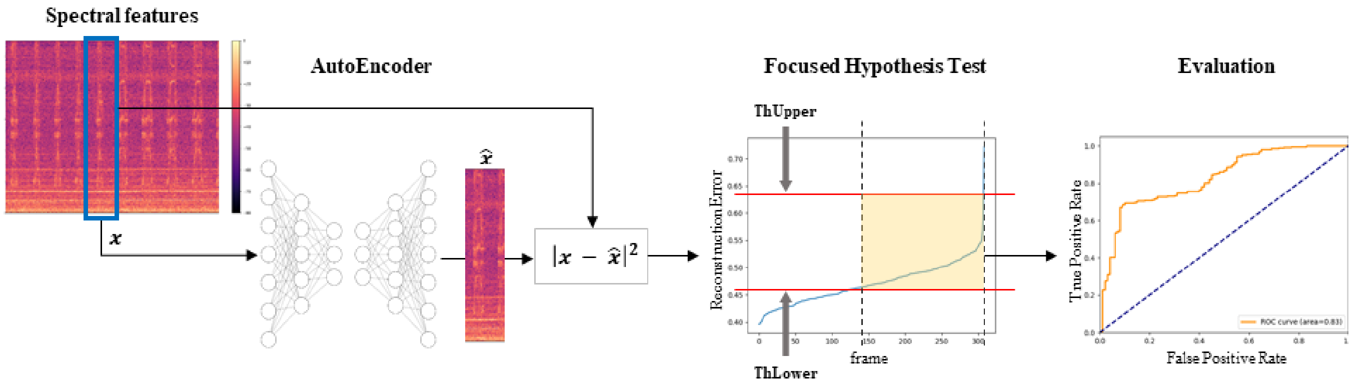

Figure 1: The block diagram of our submitted ASD system

chine sounds. By doing so, we generate normal machine sound by using latent space sampling method [4] and regard augmented normal sound samples to decrease averaged anomalous scores.

In this system, machine type-wise augmentation model has been used to construct augmented normal sound samples. The augmentation model is trained to jointly minimize Kullback-Leibler divergence (KLD) and reconstruction error (RE). KLD and RE are defined as given by

$$\begin{aligned}\mathcal{J}_{joint} &= \mathcal{J}_{KLD} + \mathcal{J}_{RE}\\ &= D(\mathcal{N}(\mathcal{E}_{gen}(\mathbf{x})|\mathbf{0},\mathbf{I})||\mathcal{N}(\mathcal{E}_{gen}(\mathbf{x})|\mu_{gen},\mathbf{\Sigma}_{gen}))\\ &\quad + E\{||\mathbf{x} - \mathcal{D}_{gen}(\mathcal{E}_{gen}(\mathbf{x}))||_2^2\} \end{aligned} \tag{3}$$

where $D(\cdot||\cdot)$ and $\mathcal{N}$ represent KLD and multivariate Gaussian distribution, respectively. Also, $\mathcal{E}_{gen}$ and $\mathcal{D}_{gen}$ are the encoder and decoder model parameters of the augmentation model. Furthermore, statistics parameters of the encoder output are calculated as μ_{gen}, $\mathbf{\Sigma}_{gen}$. These cost functions make that encoded output has a pre-defined distribution, zero-mean and unit-variance multivariate Gaussian model, and generated normal machine sound sample is more likely normal sound in the provided training set.

After training the augmentation model, acoustic features are randomly sampled from the multivariate Gaussian model $\mathcal{N}(\mathbf{0},\mathbf{I})$, then normal machine sound samples are generated from the decoder part of augmentation model. One of our submitted system uses the augmented normal machine sound samples to improve ASD accuracy.

3.2. Machine-type-wise feature selection

In the provided baseline system, AE is trained toward minimizing reconstruction error between input and reconstructed log-mel spectrogram feature vector. However, in [2], some types of machine sound have a limitations to improve ASD accuracy. To avoid limitation of ASD performance, we consider several time-frequency-domain spectral features such as linear-scale spectrogram, harmonics and percussive source separation (HPSS) [7], and median filtered spectrogram [8]. In Fig. 2, comparison result of feature vector extracted from a normal valve sound is shown. As shown in Fig. 2 (a) and (b), we found that significant pattern loss with using log-mel spectrogram feature vector. Since these dimension reduction might occur performance degradation in ASD, submitted systems are trained on linear-scale-based spectral feature vectors to overcome loss of recognizable pattern. Feature vectors are differently selected in machine type according to entire ASD accuracy. The result of ASD accuracy across different features is summarized in Table I.

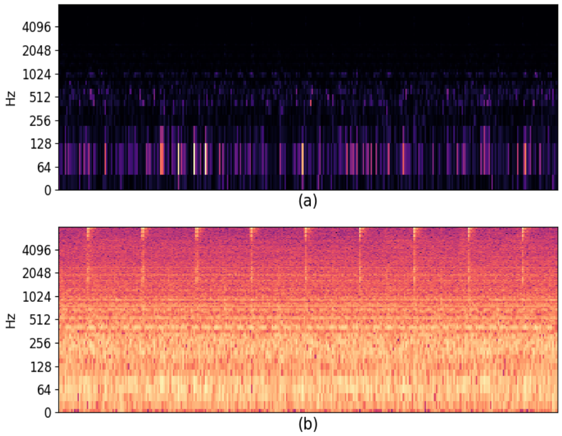

Figure 2: Comparison of spectral feature vectors extracted from normal value sound (a) log-mel spectgorgam (b) linear-scale spectrogram

3.3. Focused hypothesis test

In the provided training set, normal machine sound was recorded in the presence of factory noise and reverberation. There are none label information of normal machine sound period. After the cost function of AE model converges, baseline ASD system makes decision according to hypothesis test results $\mathcal{H}$ as follows:

$$\mathcal{H}(\mathcal{A}(\mathbf{x},\hat{\mathbf{x}})) = \begin{cases} 1(\text{Anomalous}), & \mathcal{A}(\mathbf{x},\hat{\mathbf{x}}) > \phi \\ 0(\text{normal}), & \text{otherwise} \end{cases} \tag{4}$$

where ϕ is a pre-defined threshold value. Since averaged RE in frames are considered to decide the status of machine sound in (4), averaging RE in the machine sound periods has advantage to improve entire ASD performances. However, in the unsupervised

Table 1: Performance of the baseline AE model across various spectral features

Feature vector	AUC (%)							pAUC (%)						
(dimension)	1	2	3	4	5	6	Avg	1	2	3	4	5	6	Avg
Log-mel (128)	78.77	72.53	65.83	72.89	84.76	66.28	73.51	67.58	60.43	52.45	59.99	66.53	50.98	59.66
Spectrogram (512)	78.76	71.03	64.17	**74.50**	**92.35**	**82.87**	**77.28**	63.81	60.12	51.86	**62.35**	**76.92**	56.11	**61.86**
Log-mel(128) +spectrogram(512)	69.78	**76.50**	66.17	70.20	91.97	81.42	76.01	57.68	**62.67**	52.35	62.04	75.88	**56.22**	61.14
Log-mel(128) +MF spectrogram(512)	**80.78**	74.57	**67.29**	73.14	87.22	72.16	75.86	**68.54**	61.18	**52.55**	60.83	69.62	51.50	60.70

learning scenario, we cannot precisely choose RE in the machine sound periods. To focus on RE in the normal sound periods indirectly, focused hypothesis test is adopted to our submitted systems. We found that RE of machine sound periods is lower than noise period when the machine types are valve and slider. In that case, frame energy of normal machine sounds are significantly larger than noise sounds. On the other hand, other types of machine sound show opposite pattern. From these observations, we rectify RE in frames by sorting ascending order and rejecting outliers as given by

$$\tilde{\mathcal{A}}(\mathbf{x}, \hat{\mathbf{x}}, \phi_l, \phi_u) = \begin{cases} \text{pass}, & \phi_l < \mathcal{A}_{sort}(\mathbf{x}, \hat{\mathbf{x}}) < \phi_u \\ \text{reject}, & \text{otherwise} \end{cases} \tag{5}$$

where $\mathcal{A}_{sort}(\mathbf{x}, \hat{\mathbf{x}})$ is a sorted version of $\mathcal{A}(\mathbf{x}, \hat{\mathbf{x}})$ in ascending order. Additionally, ϕ_l and ϕ_u are the threshold values for focusing RE, which are chosen empirically and differently for each machine. The concept of the focused hypothesis test is depicted in Fig. 1.

4. EXPERIMENTS AND SUBMISSIONS

We evaluated our system performances using the officially provided training set [5]-[6]. The training set consists of only normal machine sound. Each machine sound was recorded with a single microphone and sampled at 16 kHz. The recorded machine sound contains the normal machine sound as well as the factory noise signal, and label of the machine periods was not provided. To train our ASD systems, we extracted various spectral features as follows: spectrogram, HPSS, and median filtered spectrogram. The spectrogram feature was extracted with a frame size of 1024 and half overlap. Also, HPSS and median filtered feature were simply applied by using numerical python library, librosa [9]. Normal sound was generated about 30% of the entire training set. For optimal learning, the ADAM optimizer with learning rate 0.001 was set to training AE model of our ASD systems. The dimension of each spectral feature and comparison of performances across the features are summarized in Table 1 where machine classes are replaced as numbers as follows: ToyCar(1), ToyConveyor(2), Fan(3), Pump(4), Slider(5), Valve(6). According to the results in Table 1, it was confirmed that the optimal ASD performance in each spectral feature is different for each machine. In accordance with this aspect, we constructed four ASD systems for submission based on whether different features are used for each machine, whether augmentation is applied, or whether focused hypothesis test is applied. Best performance of our proposed ASD system is summarized in Table 2.

Table 2: Machine-wise performance for the best proposed method

Machine	Baseline		Best proposed	
	AUC (%)	pAUC (%)	AUC (%)	pAUC (%)
ToyCar	78.77	67.58	**82.73**	**70.35**
ToyConveyor	72.53	60.43	**76.61**	**64.04**
Fan	65.83	52.45	**70.77**	**54.50**
Pump	72.89	59.99	**76.80**	**65.56**
Slider	84.76	66.53	**94.16**	**83.97**
Valve	66.28	50.98	**89.67**	**72.85**
Average	73.51	59.66	**81.79**	**68.55**

5. CONCLUSIONS

This technical report decribes our ASD systems submitted in DCASE 2020 Task2. We applied normal sound augmentation, various spectral features, and focused hypothesis test for improving ASD performance in the noisy and reverberant environment. Ensembling these methods, the best system could achieve ASD performances, AUC and pAUC above 81% and 68%, respectively.

6. REFERENCES

[1] T. Tagawa, Y. Tadokoro, and T. Yairi, “Structured denoising autoencoder for fault detection and analysis,” in Proc. 6th Asian Conference on Machine Learning, 2015, pp. 96–111.

[2] Y. Koizumi, 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 DCASE2020 Challenge Task2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring,” 2020. [Online]. Available: arXiv:2006.05822.

[3] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, “Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma,” in Proc. 25th European Signal Processing Conference (EUSIPCO), 2017, pp. 698-702.

[4] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, “Unsupervised Detection of Anomalous Sound Based on Deep Learning and the Neyman–Pearson Lemma,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 212-224, Jan. 2019.

[5] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, “MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection,” 2019. [Online]. Available: arXiv:1909.09347.

[6] Y. Koizumi, A. Saito, H. Uematsu, N. Harada, and K. Imoto, “ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection,” in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2019, pp. 313-317.

[7] J. Driedger, M. Müller, and S. Disch, “Extending Harmonic-Percussive Separation of Audio Signals,” in Proc. the inter-

national symposium on music information retrieval (ISMIR), 2014, pp. 611-616.

[8] A. Buades, B. Coll, and J. M. Morel, “A non-local algorithm for image denoising,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), 2005, pp. 60-65.

[9] 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 Proc. the 14th python in science conference, 2015, pp. 18-25.