ANOMALOUS SOUND DETECTION SYSTEM WITH SOURCE SEPARATION MODEL-BASED FEATURE EXTRACTOR

Technical Report

Seunghyeon Shin[1], *Seokjin Lee*[1,2],

[1] School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, Republic of Korea, {sh.shin, sjlee6}@knu.ac.kr
[2] School of Electronics Engineering, Kyungpook National University, Daegu, Republic of Korea

ABSTRACT

This technical report presents an anomalous detection system developed for DCASE 2024 Task 2. Our proposed system employs a neural network to extract relevant features and calculates anomaly scores using the Mahalanobis distance with a covariance estimator. Notably, our system does not rely on any attribute information from the machines, and only minor hyperparameter adjustments are required, regardless of the machine class. These characteristics align well with the intended objectives of the task. Our approach leverages signals from machines other than the training target and trains the neural network to separate these other signals. Consequently, we can train complex neural network models effectively, even with a limited number of samples. As a result, our method achieved similar result compared to the DCASE 2024 Task 2 baseline model despite the insufficient training.

Index Terms— Anomaly detection, feature extraction, neural network, source separation,

1. INTRODUCTION

The objective of DCASE 2024 Task 2[1] is to discriminate whether an acoustic signal is normal or abnormal. This year, participants are required to develop an anomaly detection system using acoustic signals, some types of machines containing attribute information, while others do not. The anomaly detection system can use only normal condition sound clips, and each machine type has 1,000 sound clips ranging from 6 to 10 seconds in duration. Since attribute information for some types of machine is not provided, training strategies that utilize attribute information may not be effective for those types. To address the challenge of training with a limited number of sound clips, we propose an acoustic feature extraction network trained to suppress the target machine sound while preserving sounds from other machines. After training our feature extraction network, we pooled the average values in the neural network before the decoder block. From the features extracted through average pooling, we estimate an anomalous score using a covariance estimator and the Mahalanobis distance.

2. METHODOLOGY

2.1. Training strategy and anomaly score calculation

We utilized a neural network as a feature extractor. The purpose of the neural network is to extract characteristics that can distinguish whether a machine's condition is normal or abnormal. In contrast to a typical auto-encoder structure, where the neural network is trained to reconstruct the desired input signal at the output, our neural network is trained to remove the target machine signal from the input and separate the signals from other machines. In our problem, the input of the neural network $X_{t,f}$ is expressed as follows:

$$X_{t,f} = \mathcal{F}(d_c(t) + s \times n_{\bar{c}}(t)), \tag{1}$$

where $d_c(t)$ represents the target machine class c signal in time series, $n_{\bar{c}}(t)$ represents the signal from another class $\bar{c}$ signal that is not the target machine class, s is the scaling factor to match the dB of the target and other machine class signals, and the neural network input $X_{t,f}$ in the time-frequency domain is obtained by applying the short-time Fourier transform operator $\mathcal{F}$ to the sum of $d_c(t)$ and $n_{\bar{c}}(t)$.

Since we intend for our feature extractor to remove the target machine signal, we trained it to minimize the difference between the neural network's estimation output and the other class signal. We configured the training loss function $\mathcal{L}$ of the neural network as follows:

$$\begin{aligned}\mathcal{L} =& \alpha\{\frac{1}{l}\sum_{t=1}^{l}(|n_c((t) - y(t)|)\} + \beta\{\frac{1}{mn}\sum_{t=1}^{m}\sum_{f=1}^{n}(|N_{t,f}^R| - |Y_{t,f}^R|)^2\} \\ &+ \gamma\{\frac{1}{mn}\sum_{t=1}^{m}\sum_{f=1}^{n}(|N_{t,f}^I| - |Y_{t,f}^I|)^2\},\end{aligned} \tag{2}$$

where y_t and $Y_{t,f}^R$, $Y_{t,f}^I$ are the output of the neural network decoder in the time series and the real and imaginary components of the time-frequency domain, respectively. $N_{t,f}^R$ and $N_{t,f}^I$ represent the real and imaginary components of the target signal in the time-frequency domain. α, β, and γ are the hyperparameters for each difference term.

We adopt the CMGAN[2] neural network architecture as a feature extractor network. During the training procedure, the network is trained to estimate $n_{\bar{c}}(t)$ from $X_{t,f}$. After training, we explore two different approaches for average pooling. In the first approach, average pooling is applied to the output of the decoder, intermediate stage of the conformer, and conformer block at each channel, and the resulting average-pooled matrices are utilized as features for the anomaly detector. Anomaly scores are calculated from the Mahalanobis distance of the neural network output feature matrix. The covariance matrix of the feature is estimated using either maximum

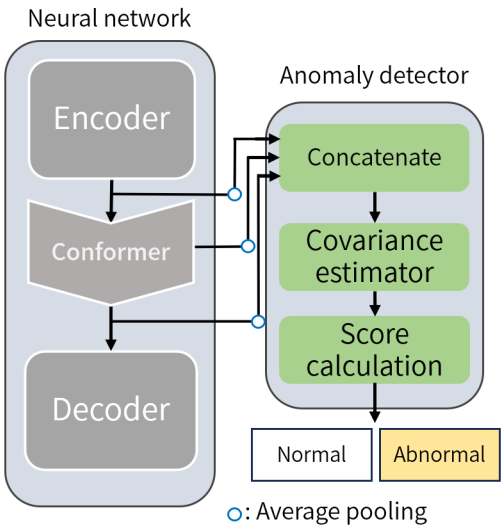

Figure 1: Anomalous detection system overview

likelihood or shrinkage estimation. In summary, we first train the neural network to separate the signal from the mixed signal of other classes and the target class signal, excluding the target class signal. After training, the average pooling is performed in multiple stages of the network or at a single point of the network. In case of a pool of multiple stages, the average pool is executed from the output of the encoder, the intermediate of the conformer, and the output of the conformer. Single-point pooling are executed from the output of the conformer. The resulting average-pooled matrices are then used as features for anomaly scoring, and in the scoring process, we employ the Mahalanobis distance with a covariance estimator. The overview of our system is shown in Fig. 1.

2.2. Experiments configure

The DCASE 2024 Task 2 utilizes two types of datasets. From the ToyADMOS2[3] and MIMII DG[4] datasets, 16 types of machine sounds are provided, each consisting of 1,000 training clips. For testing purposes, 7 machine type signals are provided, comprising 200 sound clips labeled as either normal or anomalous. Each audio clip, with a sampling rate of 16kHz, was randomly trimmed to 2 seconds and subjected to a short-time Fourier transform, using a filter length of 400 samples and an overlap of 100 samples. In the decibel matching process, the decibel levels of the other class signals were set to be 5dB lower than the target class signal. We added average pooling layers to the neural network structure of CMGAN, and average pooling was performed at each channel. Since the overall channel size of the neural network is 64, the feature vector obtained after average pooling had the same 64-dimensional size as the channel size. We perform average pooling at three locations in the neural network: the output of the encoder, the output of the second conformer block, and the output of the last conformer block, which are the inputs to the decoder. Since the network processed 2-second input segments, the 2-second average-pooled results were concatenated and used as features. The hyperparameter in the loss function α, β, and γ is used differently by the sound characteristic of the target machine. During the training, we employed the AdamW [5] optimizer and StepLR learning rate scheduler. We configured an anomalous system with four different configurations that varied the covariance estimator and the average pooling point of the neural network. We configured two options of average pooling; the first option is pooled and concatenated from the multiple stages of the neural network described previously, and the second option is pooled from the output of the conformer. In addition, two kinds of covariance estimator are used. The first is maximum likelihood estimation, and the second is shrinkage estimation. In the training procedure, with the exception of the hyperparameters α, β, and γ, all other hyperparameters and training settings remained consistent across different machine classes.

3. RESULTS AND CONCLUSIONS

The performance metric for evaluation in DCASE 2024 Task 2 is the harmonic mean of the area under the curve (AUC) and the partial area under the curve (pAUC). We compared the performance of our proposed system with the DCASE 2024 Task 2 baseline system[6], as shown in Table. 1. System IDs 1 and 2 correspond to the baseline models trained with an auto-encoder structure, where the anomalous scores were calculated using mean squared error (Baseline-MSE) and Mahalanobis distance (Baseline-Mahala), respectively. System IDs 3 and 4 utilized features obtained from three-point average pooling from the neural network, with covariance estimation performed using maximum likelihood (ML-192) and shrinkage estimation (SE-192), respectively. System IDs 5 and 6 employed features derived from single-point average pooling from the neural network, with covariance estimation using maximum likelihood (ML-64) and shrinkage estimation (SE-64), respectively. Due to the limitation of time, we did not have enough time to train the neural network sufficiently, the results of Table 1 are from the 50epoch training results. Due to time limitations, we did not have enough time to sufficiently train the neural network, and the results in Table. 1 are from the 50-epoch training. Despite insufficient training, the performance of our system is comparable to that of the baseline system, and it exhibits less variance in response to domain shifts when compared to the baseline system.

4. REFERENCES

[1] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2406.07250*, 2024.

[2] S. Abdulatif, R. Cao, and B. Yang, "Cmgan: Conformer-based metric-gan for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2477–2493, 2024.

[3] 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 Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1–5.

[4] 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

System ID	System info	Metric	bearing	fan	gearbox	slider	toycar	toytrain	valve	Official score
1	**Baseline(MSE)**	**AUC(Target)**	61.40%	55.24%	69.34%	56.01%	33.75%	46.92%	46.25%	55.35%
		AUC(Source)	62.01%	67.71%	70.40%	66.51%	66.98%	76.63%	51.07%	
		pAUC	57.58%	57.53%	55.65%	51.77%	48.77%	47.95%	52.42%	
2	**Baseline(Mahala)**	**AUC(Target)**	51.58%	42.70%	74.35%	68.11%	37.35%	39.99%	53.61%	55.01%
		AUC(Source)	54.43%	79.37%	81.82%	75.35%	63.01%	61.99%	55.69%	
		pAUC	58.82%	53.44%	55.74%	49.05%	51.04%	48.21%	51.26%	
3	**ML-192**	**AUC(Target)**	71.68%	57.28%	57.84%	61.60%	45.88%	64.88%	46.68%	55.21%
		AUC(Source)	60.12%	54.16%	64.52%	61.60%	45.40%	77.12%	47.40%	
		pAUC	56.74%	53.63%	55.47%	52.53%	48.21%	52.32%	48.26%	
4	**SC-192**	**AUC(Target)**	68.16%	66.72%	61.96%	59.68%	47.68%	63.72%	43.52%	55.41%
		AUC(Source)	65.92%	58.44%	64.68%	67.76%	38.20%	73.36%	45.60%	
		pAUC	57.21%	57.00%	54.16%	52.16%	49.42%	52.47%	48.53%	
5	**ML-64**	**AUC(Target)**	67.00%	62.68%	61.52%	59.96%	47.36%	64.44%	45.28%	55.28%
		AUC(Source)	66.08%	56.28%	65.96%	66.16%	43.44%	74.40%	45.04%	
		pAUC	56.16%	52.16%	51.63%	51.21%	49.26%	53.74%	48.26%	
6	**SC-192**	**AUC(Target)**	67.92%	65.08%	63.52%	60.52%	48.12%	63.84%	44.12%	55.16%
		AUC(Source)	65.04%	57.92%	64.24%	68.28%	37.40%	72.20%	45.08%	
		pAUC	56.47%	56.32%	51.89%	52.95%	49.68%	51.89%	48.63%	

Table 1: AUC and pAUC result of proposed model compare to baseline model

inspection for domain generalization task," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[5] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.

[6] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 191–195.