# ANOMALOUS SOUND DETECTION USING CONTRASTIVE LEARNING

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

We propose an unsupervised anomalous sound detection system for DCASE 2022 Task 2. We use self supervised contrastive learning with data augmentation as a feature extractor network. We use three kinds of data augmentation methods for contrastive learning. Then k-Nearest Neighbors are used to compute anomalous scores from extracted feature vectors. As a result, we show the detection performance of 88.58% in Area under Curve(AUC) and 74.40% in partial AUC(pAUC) with hyperparameter fixed.

***Index Terms***— Anomalous sound detection, Contrastive learning, Data augmentation

## 1. INTRODUCTION

Anomalous sound detection is the task that analysis sound to monitor machine conditions. Generally, we can obtain normal condition sound samples easily, but it is hard to obtain anomalous condition sound samples. Also, machine condition can be anomalous for many reasons and machine operating sound varies by different anomalous factors. For these reasons, the DCASE 2022 Challenge Task 2[1] is for unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques. This task requires using only normal condition samples during the training process. Also, the detector should detect anomalous sample which shift domains from the domain of training data. We propose self-supervised contrastive learning with data augmentation with conventional k-Nearest Neighbors(k-NN) anomalous detection method.

## 2. ANOMALOUS SOUND DETECTION METHOD

### 2.1. Data Preprocessing and Setup

The data of DCASE 2022 Task 2[2] consists with seven machine types and each machine type consists with six sections which is dedicated to domain shift. Two kinds of domain exists at each section. Most training data and part of target data were recorded as source and few of training data and part of test data were recorded as target domain. Each audio file is 10 seconds long with 16kHz sampling rate. Our system uses Mel spectrogram as input. We converted audio file to spectrogram using STFT with 2048 filter length and 512 hop size. And each spectrum was compressed through a Mel filter with a number of bins of 256. We fixed hyperparameter of our model overall process, all model for each machine type trained without hyperparameter change.

### 2.2. Self-supervised contrastive learning

Contrastive representation learning, which uses data augmentation as a part of architecture proposed in SimCLR[3] and showed performance of self-supervised contrastive learning can be similar compare to supervised learning. In contrastive learning process, network learns representations by minimizing agreement between data augmented from other sample in latent space and maximizing agreement between original sample and data augmented from the same sample. Our system apply ResNet-18[4] network architecture to obtain visual representation both of sample and augmented sample. After ResNet-18 network, simple projection head consisted with one hidden layer and ReLU non-linear activation function applied to visual representation. Output of projection header are used to calculate constractive loss named normalized temperature-scaled cross entropy(NT-Xent)[3] loss. NT-Xent loss function maximize agreement between augmented from same sample input and minimize agreement between augmented from different sample input. NT-Xent loss function defined as

$$l_{i,j} = -log \frac{exp(sim(z_i, z_j/\tau))}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} exp(sim(z_i, z_k/\tau))}, \quad (1)$$

where $\mathbb{1}_{[k \neq i]}$ is indicator function which is 1 when $k \neq i$, $\tau$ is temperature parameter, $(z_i, z_j)$ are data augmented from sample $i$ from batch $N$ and $sim(z_i, z_j)$ is dot product between $l_2$ normalized $(z_i, z_j)$. When data augmented from different sample, NT-Xent minimize agreement of between each sample. Negative NT-Xent loss also defined at SimCLR paper[3]. Structure of self-supervised contrastive learning model are shown in Figure 1. In Figure 1, data augmentation $x_{1,i}$ means data augmentation $i$ applied to sample $x_1$ Because of limitation of hardware, we use much lower batch size 16 compare to 4096 of SimCLR and because of small batch size, we applied Adam[5] optimizer instead of LARS optimizer[6]. As a learning rate scheduler, we use Cosine annealing warm restarts scheduler[7]. Temperature parameter $\tau$ set to 0.5.

### 2.3. Data augmentation method

In the contrastive learning process, the network learns representations from the augmented sample in latent space. Because networks learn from the augmented sample, the data augmentation method directly affects performance. So we use three kinds of data augmentation. Three data augmentation methods are illustrated in Figure 2. The first augmentation method is harmonics modification illustrated in Figure 2 (b). The harmonics modification method estimates fundamental frequency, then emphasizes odd harmonics component intensity and decreases even harmonics component intensity. The sec-

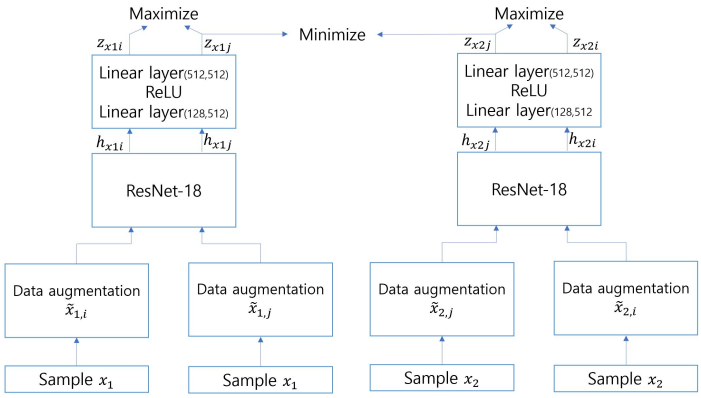


Figure 1: Self-supervised contrastive learning structure.

ond augmentation method is temporal masking illustrated in Figure 2 (c). Temporal masking masks almost 160ms duration in the time domain and the masking point is determined by the intensity at the time. The third augmentation method is F0 masking illustrated in Figure 2 (d). F0 masking augmentation method estimates the fundamental frequency of the input spectrogram and masks the estimated fundamental frequency. Three kinds of data augmentation methods are randomly applied to the input sample.

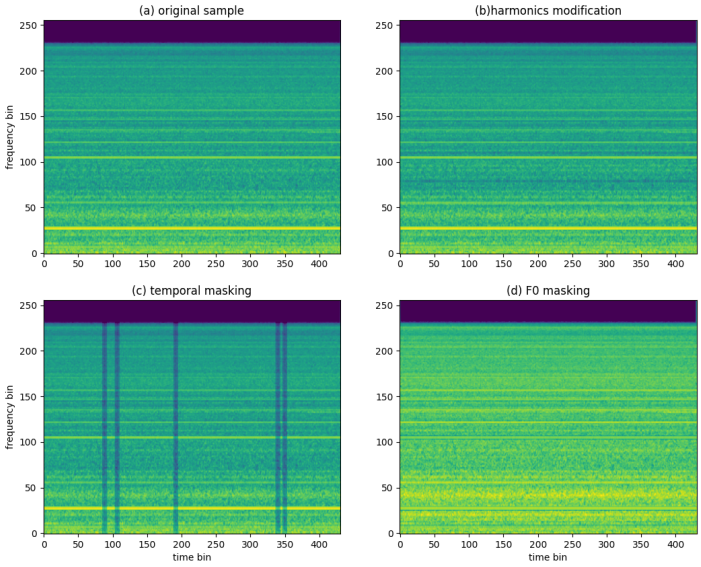


Figure 2: Augmentation illustration.

### 2.4. Anomalous detector

We use k-Nearest Neighbors(k-NN) to calculate anomalous score. Input of k-NN is 1x512 size feature vector which is output of feature extractor network. Cosine distances are used as k-NN metric and value of k set to 3.

## 3. RESULTS

In DCASE 2022 Task 2, Area Under Curve(AUC) and partial AUC(pAUC) are used as performance evaluation metrics. We compare our system with DCASE 2022 Task2 baseline system[8] in Table 1 and Table 2. AUC score per each section and domain are shown in Table 1 and pAUC scores per each section are shown in Table 2. Harmonic mean of AUC scored 88.58% compare to the score of baseline 52.07% and 56.09% and harmonic mean of pAUC scored 74.40% compared to scores of baseline of 53.73% and 55.65%. The largest AUC score fluctuation between source and target domain is 7.28% for ToyTrain machine type and the largest pAUC score fluctuation is 7.91% for ToyCar machine type. Score fluctuation shows proposed model extracts domain generalized features.

| **dataset split** | | | **baseline system** | | **proposed system** |
|---|---|---|---|---|---|
| | | | **autoencoder** | **MobileNetV2** | |
| machine type | section | domain | AUC | AUC | AUC |
| ToyCar | 0 | source | 86.42% | 47.40% | 90.58% |
| ToyCar | 1 | source | 89.85% | 62.02% | 86.60% |
| ToyCar | 2 | source | 98.84% | 74.19% | 83.72% |
| ToyCar | 0 | target | 41.48% | 56.40% | 89.64% |
| ToyCar | 1 | target | 41.93% | 56.38% | 90.84% |
| ToyCar | 2 | target | 26.50% | 45.64% | 95.40% |
| ToyCar | harmonic mean | | 50.27% | 55.31% | 89.31% |
| ToyTrain | 0 | source | 67.54% | 46.02% | 73.84% |
| ToyTrain | 1 | source | 79.32% | 71.96% | 71.62% |
| ToyTrain | 2 | source | 84.08% | 63.23% | 83.92% |
| ToyTrain | 0 | target | 33.68% | 49.41% | 87.78% |
| ToyTrain | 1 | target | 29.87% | 45.14% | 80.74% |
| ToyTrain | 2 | target | 15.52% | 44.34% | 81.96% |
| ToyTrain | harmonic mean | | 35.76% | 50.95% | 79.58% |
| bearing | 0 | source | 57.48% | 67.85% | 86.16% |
| bearing | 1 | source | 71.03% | 59.67% | 82.04% |
| bearing | 2 | source | 42.34$ | 61.71% | 96.28% |
| bearing | 0 | target | 63.07% | 60.17% | 87.86% |
| bearing | 1 | target | 61.04% | 64.65% | 79.54% |
| bearing | 2 | target | 52.91% | 60.55% | 95.10% |
| bearing | harmonic mean | | 56.33% | 60.26% | 87.40% |
| fan | 0 | source | 84.69% | 71.07% | 96.70% |
| fan | 1 | source | 71.69% | 76.26% | 83.52% |
| fan | 2 | source | 80.54% | 67.29% | 91.46% |
| fan | 0 | target | 39.35% | 62.13% | 96.80% |
| fan | 1 | target | 44.74% | 35.12% | 88.62% |
| fan | 2 | target | 63.49% | 58.02% | 88.22% |
| fan | harmonic mean | | 58.96% | 57.35% | 90.64% |
| gearbox | 0 | source | 64.63% | 63.54% | 94.60% |
| gearbox | 1 | source | 67.66% | 66.68% | 92.44% |
| gearbox | 2 | source | 75.38% | 80.87% | 92.96% |
| gearbox | 0 | target | 64.79% | 67.02% | 92.44% |
| gearbox | 1 | target | 58.12% | 66.96% | 87.44% |
| gearbox | 2 | target | 65.57% | 43.15% | 89.04% |
| gearbox | harmonic mean | | 65.63% | 62.02% | 89.59% |
| slider | 0 | source | 81.92% | 87.15% | 91.50% |
| slider | 1 | source | 67.85% | 49.66% | 84.28% |
| slider | 2 | source | 86.66% | 72.70% | 82.68% |
| slider | 0 | target | 58.04% | 80.77% | 87.82% |
| slider | 1 | target | 50.3% | 32.07% | 86.48% |
| slider | 2 | target | 38.78% | 32.94% | 96.88% |
| slider | harmonic mean | | 59.16% | 48.19% | 88.03% |
| valve | 0 | source | 54.24% | 75.26% | 86.16% |
| valve | 1 | source | 50.45% | 54.78% | 82.04% |
| valve | 2 | source | 51.56% | 76.26% | 96.28% |
| valve | 0 | target | 52.73% | 43.60% | 87.86% |
| valve | 1 | target | 53.01% | 60.43% | 79.54% |
| valve | 2 | target | 43.84% | 78.74% | 95.10% |
| valve | harmonic mean | | 50.70% | 61.76% | 87.03% |
| all | harmonic mean | | 52.07% | 56.09% | 88.58% |

Table 1: AUC score result

| dataset split | | baseline system | | proposed system |
|---|---|---|---|---|
| | | **Autoencoder** | **MobileNetV2** | |
| machine type | section | pAUC | pAUC | pAUC |
| Toycar | 0 | 51.31% | 49.96% | 76.02% |
| Toycar | 1 | 54.08% | 50.92% | 73.78% |
| Toycar | 2 | 52.96% | 56.51% | 75.53% |
| Toycar | harmonic mean | 52.74% | 52.27% | 75.09% |
| ToyTrain | 0 | 52.72% | 50.25% | 67.01% |
| ToyTrain | 1 | 50.64% | 52.97% | 59.53% |
| ToyTrain | 2 | 48.33% | 51.54% | 64.58% |
| ToyTrain | harmonic mean | 50.48% | 51.52% | 63.55% |
| bearing | 0 | 51.49% | 54.41% | 69.73% |
| bearing | 1 | 55.85% | 55.09% | 71.51% |
| bearing | 2 | 49.18% | 64.18% | 85.74% |
| bearing | harmonic mean | 51.98% | 57.14% | 75.02% |
| fan | 0 | 59.95% | 55.40% | 89.42% |
| fan | 1 | 51.12% | 52.14% | 71.19% |
| fan | 2 | 62.88% | 65.14% | 77.66% |
| fan | harmonic mean | 57.52% | 56.9% | 78.72% |
| gearbox | 0 | 60.93% | 62.12% | 81.71% |
| gearbox | 1 | 53.74% | 56.85% | 79.40% |
| gearbox | 2 | 61.51% | 50.62% | 81.69% |
| gearbox | harmonic mean | 58.49% | 56.03% | 80.92% |
| slider | 0 | 61.65% | 71.57% | 76.18% |
| slider | 1 | 53.06% | 48.21% | 71.54% |
| slider | 2 | 53.44% | 49.69% | 77.76% |
| slider | harmonic mean | 55.78% | 54.67% | 75.06% |
| valve | 0 | 52.15% | 55.37% | 69.73% |
| valve | 1 | 49.78% | 54.69% | 71.51% |
| valve | 2 | 49.24% | 85.74% | 85.74% |
| valve | harmonic mean | 50.36% | 62.42% | 75.02% |
| all | harmonic mean | 53.73% | 55.65% | 74.40% |

Table 2: pAUC score result

## 4. CONCLUSIONS

In this work, self-supervised contrastive learning based anomalous sound detection system are proposed. In the system, self-supervised contrastive learning network extracting features from data augmentation applied sample. Three kinds of data augmentation methods are used to extract domain generalized feature. Extracted features are used to calculate anomalous score by k-NN algorithm. As a result, proposed system scored AUC score 88.58% and pAUC score 74.40% with small performance fluctuations when domains are changed.

## 5. 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] 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.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[5] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[6] Y. You, I. Gitman, and B. Ginsburg, "Large batch training of convolutional networks." [Online]. Available: https://arxiv.org/abs/1708.03888

[7] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," 2016. [Online]. Available: https://arxiv.org/abs/1608.03983

[8] 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.