# DCASE2020 TASK2 SELF-SUPERVISED LEARNING SOLUTION

Technical Report

*Zero Shinmura*

Nagano, Japan
shinmura_0@yahoo.co.jp

## ABSTRACT

The detection of anomalies by sound is very useful. Because, unlike images, there is no need to worry about adjusting the light or shielding. We propose anomaly sound detection method using self-supervised learning with deep metric learning. Our approach is fast because of using MobileNet V2. And our approach was good at non-stationary sounds, achieving an AUC of 0.9 or higher for most of non-stationary sounds.

***Index Terms***— Self-supervised learning, deep metric learning

## 1. INTRODUCTION

The detection of anomalies with cameras has become very accurate due to the development of deep learning. However, cameras may not work well due to waterproofing issues, shielding, or lack of light. In this respect, the abnormal sound detection is very useful. It easily clears up shielding and waterproofing issues, and it works in the dark. In that sense, this competition (DCASE2020 Task2) was quite practical and fun for detecting abnormal sounds. Thanks to the organizers.

Deep learning has also been shown to be effective in abnormal sound detection. Sound data is usually converted into a (Mel-) spectrogram for processing. This transformation will enable the sound to be represented in 2 dimensions, and deep learning can make sound data easier to handle. And we used this style.

However, there are other methods that do not use the (Mel-) spectrogram. Some studies have used raw sound waveforms for environmental sound recognition. EnvNet[1] is a representative one, which extracts features using a Convolutional Neural Network(CNN) on raw waveforms. EnvNet alone performed as well as the previous work (using Mel-spectrogram). However, the accuracy was greatly improved by combining EnvNet and the previous work. We also wanted to make a model that mixed raw waveforms and Mel-spectrogram, but we couldn't do that because of the PC's memory. Instead, we trained a model using spectrogram. And we didn't use raw waveforms.

## 2. RELATED WORKS

A common method of abnormal sound detection is an autoencoder. This method was also used at baseline[12][13][14] and with this showed good scores. But we think autoencoder is the problem method is that it is sensitivity to noise and difficult to find small abnormalities. So in this competition we used self-supervised learning (SSL).

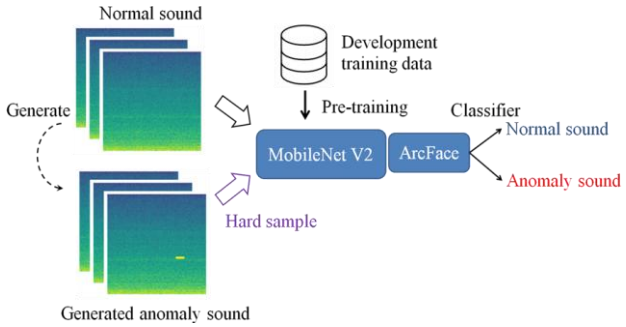


Figure 1: The architecture of proposed method.

One such study of SSL for anomaly detection is using geometric transformations [2]. This method, generate transformed images of normal images by geometric transformations. In this method, anomaly score is class score of geometric transformations. And this method was increasing AUC score greatly than the previous work. We used SSL but didn't use geometric transformations. Our SSL is generating anomaly data from normal data. And a classification model is trained with normal data and generated anomaly data.

## 3. PROPOSED METHOD

In this section, we describe our approach in this competition. The proposed method is shown in Fig.1.

### 3.1. Self-Supervised Learning (SSL)

The basis of our approach is SSL. Our approach artificially generate anomaly data from normal data (spectrograms form).And we made CNN train the binary classification (normal or anomaly) with generated data. In order to generate anomaly data from normal data, we show examples of test data in Fig.2. Valve, slider and toy car is non-stationary sounds. On the other hand fan, pump and toy conveyor is stationary sounds. In anomaly sounds of stationary sounds, short sounds occur.

The anomaly data we generated is shown in Fig.3.White noise adds noise across all frequencies. On the other hand the pink noise adds noise to the lower frequencies with emphasis. Many industrial machines have characteristic sounds at low frequencies, and pink noise was found to be effective in disrupt-

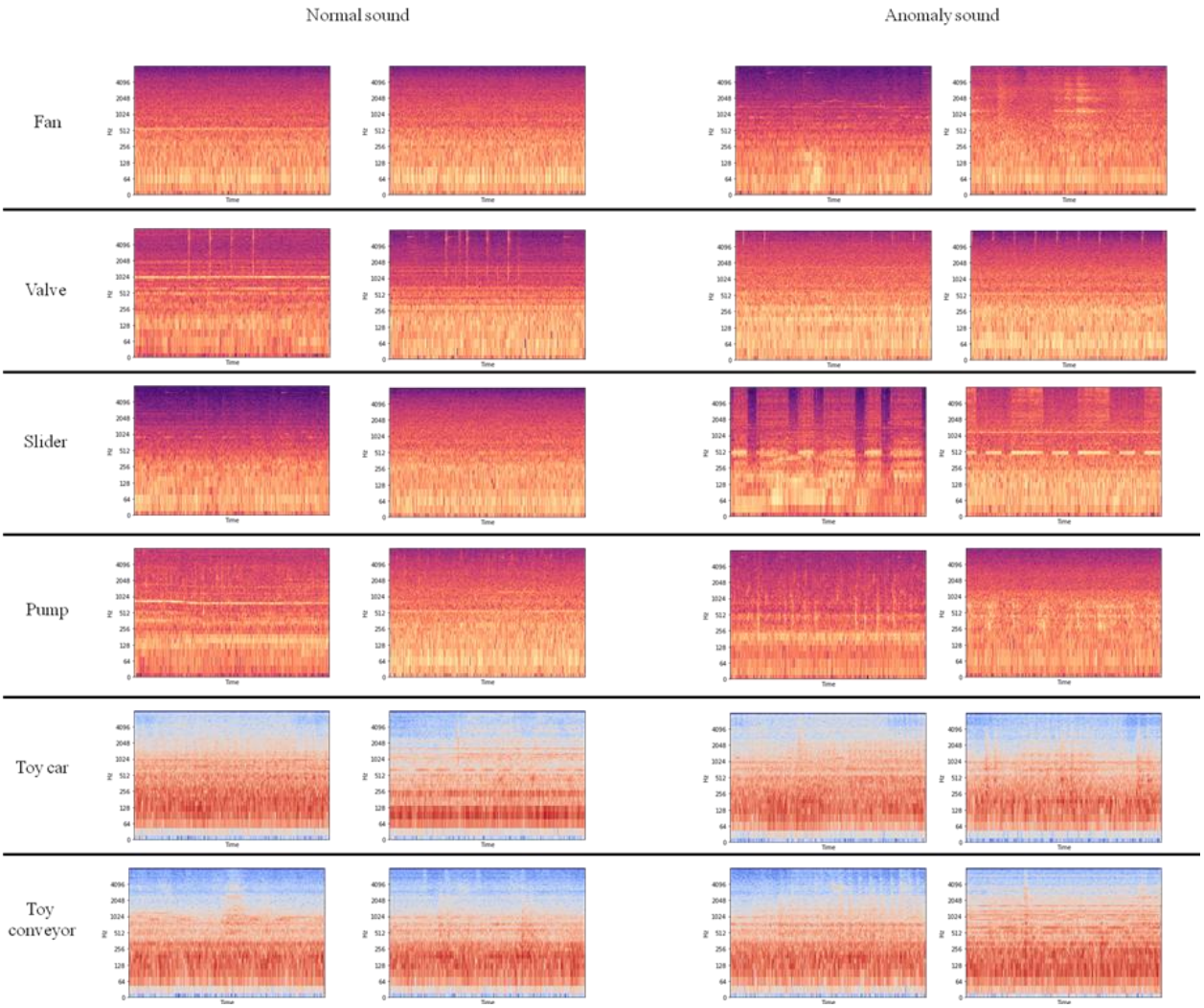


Figure 2: Examples of test data

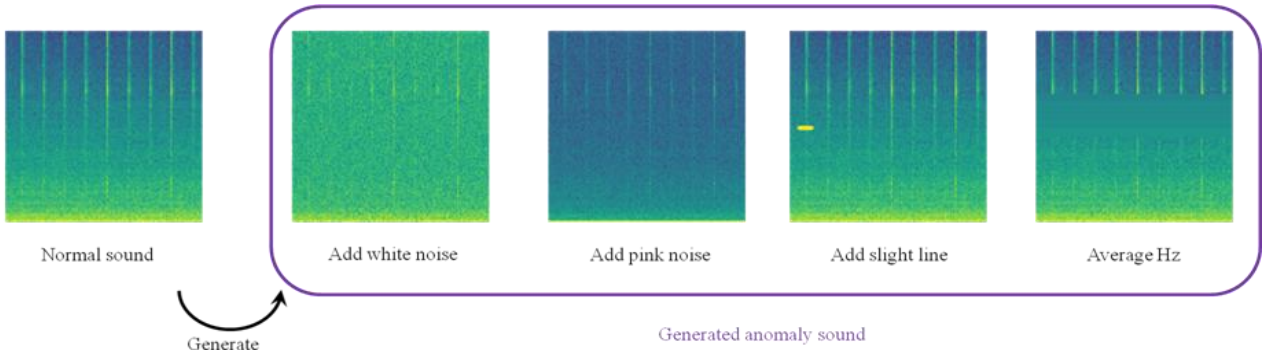


Figure 3: Generated anomaly sound from normal sound

ing these characteristic sounds. The slight line assumes following abnormal sounds. Bearings and other rotating parts may periodically occur high-frequency abnormal sounds when the grease runs out. The slight line assumes such an abnormal sound. Average Hz is averaged in time directions. This disturbs the sound time rules.

### 3.2. ArcFace

In anomaly detection, using deep metric learning improves the accuracy [3]. First we tried to use Circle Loss [5] as the latest deep metric learning method, but the score was not good enough in this competition, so we used ArcFace[4].

Table 1: Experimental result (ROC-AUC)

| | fan | | | | valve | | | | slider | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | 0 | 2 | 4 | 6 | 0 | 2 | 4 | 6 | 0 | 2 | 4 | 6 |
| Baseline | 0.540 | 0.722 | 0.622 | 0.723 | 0.687 | 0.674 | 0.744 | 0.553 | 0.963 | 0.788 | 0.946 | 0.702 |
| IDNN | – | – | – | – | 0.917 | 0.928 | 0.923 | 0.786 | 0.982 | 0.859 | 0.978 | 0.836 |
| **Ours** | 0.642 | 0.948 | 0.775 | 0.691 | 0.996 | 0.999 | 0.959 | 0.808 | 0.999 | 0.943 | 0.999 | 0.951 |

| | pump | | | | ToyCar | | | | ToyConveyor | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | 0 | 2 | 4 | 6 | 1 | 2 | 3 | 4 | 1 | 2 | 3 |
| Baseline | 0.671 | 0.609 | 0.887 | 0.735 | 0.828 | 0.866 | 0.639 | 0.869 | 0.793 | 0.642 | 0.745 |
| IDNN | – | – | – | – | 0.776 | 0.792 | 0.578 | 0.624 | – | – | – |
| **Ours** | 0.899 | 0.873 | 0.999 | 0.787 | 0.823 | 0.899 | 0.919 | 0.995 | 0.748 | 0.628 | 0.724 |

### 3.3. Hard sample

Deep metric learning is a powerful classifier. However, it is known that the classification accuracy is changed by the way data is given. Some studies [6][7] have reported that the accuracy increases when given data are difficult to classify. This method was also adopted in our approach.

Our method is simple.1) Take a batch of normal (trainging) data. 2) Generate anomaly data from batch normal data. This size is 5times of batch normal data. 3) Select the generated data that is difficult to classify (the lower accuracy score of the anomaly class). The size of selected data is batch size. 4) Train a model with normal data and the selected data. 5) repeat with every batch.

### 3.4. Pre-training

In image recognition, it is known that accuracy increases when using a pre-trained model. It was also reported that the accuracy of sound recognition was greatly improved by using the pre-training [8][9], and this was applied to our approach.

The rules of DCASE2020 allowed to use a few pre-trained models (such as VGGish). However, when we used VGGish, the score was low, so we prepared pre-training model using only the development training data (not test data) published in Task 2. We selected 18 different types data (selected 3 types sound per a machine) and we converted them into spectrogram. The base architecture was MobileNetV2 and we trained this model (random initial weights) with SoftMax Cross entropy classification problems (18 classes) using Mixup[10]. The accuracy was 98%.

### 3.5. Anomaly score

In order to get anomaly score, we removed the ArcFace layer. Then output of model is Global Average Pooling (GAP) of MobileNet V2. In test phase, the minimum of cosine similarity between all training sound (GAP outputs) and test data was used as the abnormal score.

In order to stabilize the anomaly score, we performed an ensemble of 10 models. The ensemble method is simple and only takes the average of the anomaly scores of the 10 models.

## 4. EXPERIMENT

In this section, we evaluated our approach with development dataset. We used training data for training a model. And we used test data to evaluate.

Table 2: Effectiveness of generated anomaly sound (Fan ID=2)

| White noise | Pink noise | Slight line | Average Hz | AUC |
|---|---|---|---|---|
| ✓ | | | | 0.778 |
| ✓ | ✓ | | | 0.873 |
| ✓ | ✓ | ✓ | | 0.840 |
| ✓ | ✓ | ✓ | ✓ | 0.948 |

### 4.1. Setting

We used Librosa to convert wave file to spectrogram. Then n_fft=512, hop_length=256, and we used amplitude_to_db. Input size of MobileNet V2 was 224x224. Channel is 1. Therefore we converted spectrogram to Image (224x224) with OpenCV. And Batch size was 64 for pre-training and 32 for SSL. Preprocess was Normalization (i.e., for all data, subtracted by the mean and divided by the standard deviation). The optimization method used Adam (lr=0.0001) for SSL. Epoch was 7. The parameters of ArcFace was m=30, s=0.05.

IDNN [11] is a powerful method for detecting abnormalities in non-stationary sounds. Especially, when the sound occur periodically, the abnormality detection ability is good. We compared IDNN and our approach. The optimization method of IDNN was Adam (default). And batch size was 128. Loss function was “mean squared error”. Epoch was 100. The architecture of IDNN was 8 layers CNN autoencoder(channel filters 32-64-128-256-256-128-64-32). Activation was ReLU.

### 4.2. Result

The result was table 1. Baseline was reprinted from Github. Our approach outperformed other method for most of the data. In particular, our approach scored very high for non-stationary sounds (valve, slider, toy car). IDNN showed a good score on the valve and slider, but the score did not increase on the toy car. This is because the period of the valve and slider is constant and the waveform is clean, whereas the toy car has a partly disturbed period and the waveform can be broken.

## 5. ABLATION STUDY

To discuss our approach, we conduct an Ablation study. The result is a table 2 and 3. According to Table 2, average Hz and pink noise is effective. This is because fan is characterized by low frequencies and the pink noise disturbs this. In the anomaly sound of fan, average Hz was also effective because the abnorm-

Table 3: Effectiveness of generated anomaly sound (valve ID=4)

| White noise | Pink noise | Slight line | Average Hz | AUC |
|---|---|---|---|---|
| ✓ | | | | 0.750 |
| ✓ | ✓ | | | 0.782 |
| ✓ | ✓ | ✓ | | 0.765 |
| ✓ | ✓ | ✓ | ✓ | 0.959 |

al sound was instantaneous. According to Table 3, average Hz has the greatest effect. This had a similar effect for all non-stationary sounds (slider, toy car), not just the valves. Average Hz has the effect of disturbing a sound with a period in time. Since many of the non-stationary sounds have a period, average Hz was effective. And thanks to average Hz, the score increased significantly in non-stationary sounds.

In table 4, we evaluate effect of pre-training model and hard sample. Surprisingly, hard sample had the greatest effect. The use of hard sample in deep metric learning has been reported to improve accuracy in many cases, but it was surprising. The reason for this is that the generated anomaly data, which is difficult to classify, is different for each machine sounds. For example, valve has a strong sound at regular intervals, so it's easy to recognize the sound generated by average Hz (disturbing interval). However, valve sound are characteristic at high frequencies, the sounds generated by pink noise (which disturbs low frequencies) are likely to be difficult to recognize. Then most of hard sample selected the sounds generated by pink noise (not average Hz) in valve. In toy car, hard sample is selected sounds different from valve. As you can see, each mechanical sound has its own personality, and there are different sounds that are difficult to classify. Hard sample provides anomaly sounds appropriate for each machine sound in order to train the good model.

Finally, we present an approach that we failed. First, We tried to remove the normal sound by NMF (Nonnegative Matrix Factorization) in training data, then I tried to determine if it was normal or not by the remaining sound, but it failed. Second, there were many machines that had a characteristic sound in the low-frequency component, so they were separated into 5 channels in the frequency direction. However, they were not equally spaced, and were separated so that the low frequencies were more emphasized, but this did not have much effect. Third, we tried to ensemble our approach with IDNN, but unlike supervised learning, the anomaly detection ensemble was difficult and did not score well.

## 6. CONCLUSION

In this paper, we presented self-supervised learning (SSL) solution for detection anomaly sound. The SSL scores were particularly high for non-stationary sounds (valve, slider, toy car). This is due in large part to the average Hz technique. But in stationary sound (fan, toy conveyor), the score didn't increase. As a future work, I would like to study SSL techniques that can achieve high scores even with stationary sounds.

Table 4: Effectiveness of pre-training and hard sample (Fan ID=2)

| SSL | Pre-training | Hard sample | AUC |
|---|---|---|---|
| ✓ | | | 0.511 |
| ✓ | ✓ | | 0.750 |
| ✓ | ✓ | ✓ | 0.948 |

## 7. REFERENCES


[1] Yuji Tokozume and Tatsuya Harada. Learning environmental sounds with end-to-end convolutional neural network. In ICASSP, 2017.

[2] Izhak Golan and Ran El-Yaniv. Deep Anomaly Detection Using Geometric Transformations. arXiv preprint arXiv: 1805.10917, 2018.

[3] Liron Bergman and Yedid Hoshen. Classification-Based Anomaly Detection for General Data. arXiv preprint arXiv: 2005.02359, 2020.

[4] Jiankang Deng, Jia Guo, Niannan Xue and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. arXiv preprint arXiv:1801.07698, 2018.

[5] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang and Yichen Wei. Circle Loss: A Unified Perspective of Pair Similarity Optimization. arXiv preprint arXiv:2002.10857, 2020.

[6] Byungsoo Ko and Geonmo Gu. Embedding Expansion: Augmentation in Embedding Space for Deep Metric Learning. arXiv preprint arXiv: 2003.02546, 2020.

[7] Wenzhao Zheng, Zhaodong Chen, Jiwen Lu and Jie Zhou. Hardness-Aware Deep Metric Learning. arXiv preprint arXiv:1903.05503, 2019.

[8] Andrey Guzhov, Federico Raue, Jörn Hees and Andreas Dengel. ESResNet: Environmental Sound Classification Based on Visual Domain Models. ArXiv:2004.07301,2020.

[9] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang and Mark D. Plumbley.PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. arXiv preprint arXiv:1912.10211, 2019.

[10] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. arXiv preprint arXiv:1710.09412, 2017.

[11] Kaori Suefusa, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo and Yohei Kawaguchi. Anomalous Sound Detection Based on Interpolation Deep Neural Network. ICASSP2020, 2020.

[12] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Noboru Harada, and Keisuke Imoto. ToyADMOS: a dataset of miniature-machine operating sounds for anomalous sound detection. In Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 308–312. 2019.

[13] Harsh Purohit, Ryo Tanabe, Takeshi Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi. MIMII Dataset: sound dataset for malfunctioning industrial machine investigation and inspection. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), 209–213. 2019.

[14] Yuma Koizumi, Yohei Kawaguchi, Keisuke Imoto, Toshiki Nakamura, Yuki Nikaido, Ryo Tanabe, Harsh Purohit, Kaori Suefusa, Takashi Endo, Masahiro Yasuda, and Noboru Harada. Description and discussion on DCASE2020 challenge task2: unsupervised anomalous sound detection for machine condition monitoring. In arXiv e-prints: 2006.05822. 2020.