UNSUPERVISED ANOMALOUS SOUND DETECTION USING DENOISING-DETECTION SYSTEM UNDER DOMAIN SHIFTED CONDITIONS

Technical Report

Chenxu Zhang[1]*, Yao Yao*[1]*, Rui Qiu*[1]*, Shengchen Li*[2]*, Xi Shao*[1]

[1] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China, {1018010429, 1019010528, 1220013123, shaoxi}@njupt.edu.cn
[2] School of Advanced Technology, Xi'an Jiaotong-liverpool University, Suzhou, China, Shengchen.Li@xjtlu.edu.cn

ABSTRACT

The DCASE2021 Challenge Task2 is to develop an unsupervised detection system of anomalous sounds for seven types of machines under domain shifted conditions. A common challenge in the detection of anomalous sounds for machine is to identify the diversity of malfunctioning sounds and the scarcity of malfunctioning sounds samples between normal and anomalous condition. In this paper, an unsupervised denoising-detection system is proposed to perform this task by: (1) removing noise in each recording to obtain signal that is more related to this task; (2) training an overfitting model by leveraging the information of sections in each machine type. The experimental evaluation demonstrates that the proposed system outperforms the provided baseline system across majority of machine types in both source domain and target domain.

Index Terms— DCASE2021 Task2, Unsupervised anomalous sound detection, Domain shifted conditions, Denoising, Overfitting

1. INTRODUCTION

Unsupervised anomalous detection [1, 2] detects anomalous samples under the condition that only normal samples have been provided during training phase. Such problems interests both academic and industry, and has a wide range of applications. In this work, we focus on detecting anomalous sounds for machine condition monitoring.

The anomalous sound detection is considered as an outlier detection problem [3]. It first models the normal sound through a variety of methods such as neural networks. The deviation between the model and the observed sound is then calculated, usually called an anomaly score. When the anomaly score is higher than a predetermined threshold, the observed sound is recognized as anomalous.

However, in the case of real-world factories, the sounds recognized as anomalous by the above method are not always the real anomalous that we concern about. It may be caused by the differences in operating speed, machine load, environmental noise, etc. Therefore, we also aim to solve the problem of normal sounds being incorrectly judged as anomalous due to changes within the normal conditions.

In this work, all recordings include both the sound of a machine as well as the environmental sounds of factoriy. After removing the effects of environmental noise, the sounds caused by malfunction will become obvious. In addition to environmental noise, the changes within the normal conditions are also regarded as noise. Although the conditions have changed, normal recording is still normal recording. These changes are useless for detecting abnormalities, that is, they do not relevant to our goal, so they can be regarded as noise. We first train a denoising network by leveraging AudioSet [4] to remove noise for all types of machines to obtain signal that is more related to this task. Then, an overfitting model based on MobileNetV2 [5] is trained for each type of machine by leveraging the information of section. The output difference between normal and abnormal will be expand due to the overfitting of normal data. Section 2 describes techniques of our system. The experimental results are shown in Section 3. Finally, our conclusions are provided in Section 4.

2. PROPOSED SYSTEM

An overview of the proposed system which is separated into pretrain, training and test phases, as shown in Figure 1. The procedure of the proposed method is described in detail in the following sections.

2.1. Audio processing and preparation

2.1.1. Audio denoising

A denoising network based on Deep Xi [6, 7] is used in the preprocessing for removing noise from the original samples and reducing the effect of conditions changes by treating it as a type of noise.

All recordings in dataset include both the sound of a machine and its associated equipment as well as environmental sounds. The conditions of source and target domains differ in terms of operating speed, machine load, viscosity, heating temperature, environmental noise, SNR, etc. In this paper, recording is regarded as a collection of useful signal and noise, where noise includes the environmental sounds and the changes within the normal conditions. Thus the goal of the denoising network is to obtain signal that is more related to this task.

In order to verify the effectiveness of denoising, a speech denoising network leveraging AudioSet is used for all types of machines rather than doing denoising works for machine types or sections separately. The noisy audio is first transformed to the frequency domain, and the statistical characteristics of the noise spectrum are obtained. Then the noise spectrum is modified according to the statistical characteristics to form a speech enhancement spec-

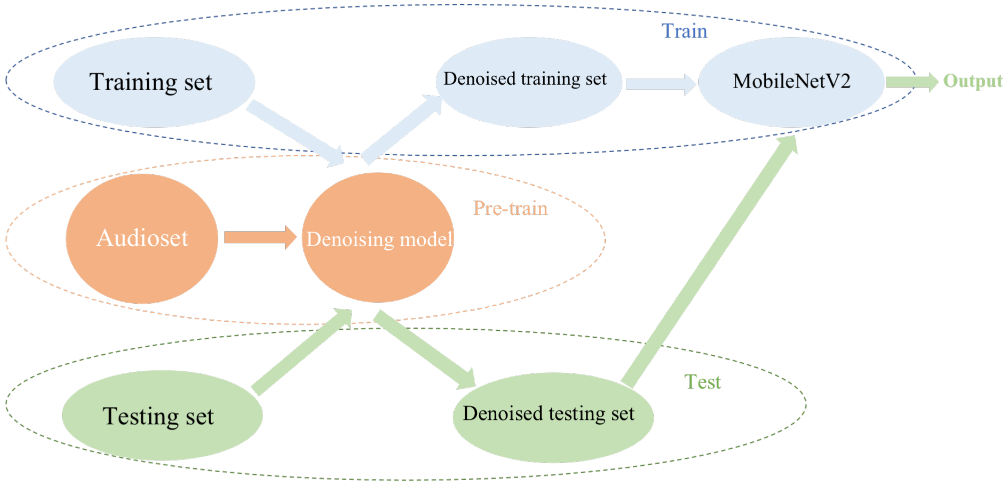

Figure 1: System overview.

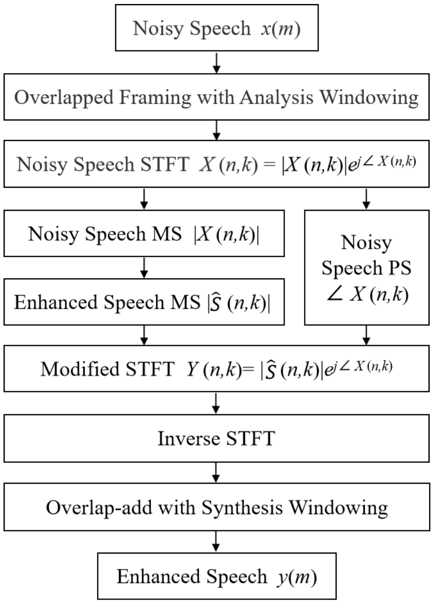
Noisy Speech $x(m)$
Overlapped Framing with Analysis Windowing
Noisy Speech STFT $X(n,k) = |X(n,k)|e^{j\angle X(n,k)}$
Noisy Speech MS $|X(n,k)|$
Enhanced Speech MS $|\hat{S}(n,k)|$
Noisy Speech PS $\angle X(n,k)$
Modified STFT $Y(n,k)= |\hat{S}(n,k)|e^{j\angle X(n,k)}$
Inverse STFT
Overlap-add with Synthesis Windowing
Enhanced Speech $y(m)$

Figure 2: The algorithm of denoising.

trum, and finally transformed back to the time domain. The algorithm is shown in Figure 2.

2.1.2. *Audio preparation*

For each machine type, linear combination of different section is used to generate additional data.

There are 3 sections in development dataset, which are section00, section01 and section02. Each section consists of data from source domain and target domain. In each section, the changes within the normal conditions are different from others. The additional data is a mixture of these section as shown in Table 1.

The additional data is similar but not identical to the original data. Hence the boundary created by the proposed system for target section will be more accurate in order to distinguish these section.

Table 1: Generation of additional data.

Section	Combination
03	00 & 01
04	00 & 02
05	01 & 02

2.2. Feature extraction

Spectral features of auditory signals are used in our system. For each 10s audio, setting the length of sample window to 64ms, hop size to 32ms, number of filters to 128 and maximum frequency to 8000Hz to get Log-mel-spectrogram feature [8]. Then, 64 consecutive frames are used as input vector. Finally, for each audio, a feature matrix whose shape is 251 × 64 × 128 × 1 is obtained.

2.3. Classifier

This part, we follow the DCASE2021 MobileNetV2 baseline but to train an overfitting model for each type of machine by leveraging the information of section.

The learning task is to create classification boundary for each section. It identifies from which section the observed signal was generated. In other words, it outputs the softmax value that is the predicted probability for each section. Due to the overfitting of the normal data, in test phase, the output of abnormal data will have a large difference with that of normal one.

The off-the-shelf Keras implementation of MobileNetV2 is used with the width multiplier parameter set to 0.5. The loss function is categorical cross-entropy and the optimization algorithm is adam with 10^{-5} learning rate. The batch size is 32 with 50 epochs, the split percentage of validation is 0.1 after data shuffle.

2.4. Outlier Detection

In this work, the anomaly score is calculated as the averaged negative logit of the predicted probabilities for the correct section, which can be described as:

$$A_\theta\left(X\right) = \frac{1}{B}\sum_{b=1}^{B} log\left\{\frac{1-p_\theta(\varphi_{t(b)})}{p_\theta(\varphi_{t(b)})}\right\}, \qquad (1)$$

where B is the num of frames, $t(b)$ is the beginning frame index of the b-th image, φ is the acoustic feature, and p_θ is the softmax output by MobileNetV2 for the correct section.

To determine the anomaly detection threshold, assuming A_θ follows the gamma distribution. The parameters of the gamma distribution are estimated from the histogram of A_θ, and the anomaly detection threshold is determined as the 90th percentile of the gamma distribution. If A_θ for each test clip is greater than this threshold, the clip is judged to be abnormal; if it is smaller, it is judged to be normal.

3. EXPERIMENTAL EVALUATION

3.1. Dataset

We evaluate the proposed system on the DCASE2021 Challenge Task2 [9] development dataset. The dataset comprises parts of MIMII DUE [10] and ToyADMOS2 [11], containing the normal and abnormal sounds of seven real machines: Fan, Gearbox, Pump, Slide rail, ToyCar, ToyTrain and Valve. Each recording is single-channel, 10-second audio that includes both the sounds of a machine and its associated equipment as well as environmental sounds. There are three sections for each machine type, and each section is a complete set of training and test data. For each section, this dataset provides (i) around 1000 clips of normal sounds in a source domain for training; (ii) only three clips of normal sounds in a target domain for training; (iii) around 100 clips each of normal and anomalous sounds in the source domain for the test, and (iv) around 100 clips each of normal and anomalous sounds in the target domain for the test.

3.2. Evaluation metrics

To evaluate the performance of our method, the anomaly scores are translated into AUC value and pAUC value. AUC [12] is defined as the area enclosed by the coordinate axis under the ROC (Receiver Operating Characteristic) curve. pAUC is calculated as the AUC over a low false-positive-rate (FPR) range $[0, p]$. In this task, $p = 0.1$. The AUC and pAUC for each machine type, section, and domain are defined as:

$$\mathrm{AUC}_{m,m,d} = \frac{1}{N_- N_+}\sum_{i=1}^{N_-}\sum_{j=1}^{N_+}\mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \qquad (2)$$

$$\mathrm{pAUC}_{m,m,d} = \frac{1}{\lfloor pN_- \rfloor N_+}\sum_{i=1}^{\lfloor pN_- \rfloor}\sum_{j=1}^{N_+}\mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \quad (3)$$

where m represents the index of a machine type, n represents the index of a section, $d = \{\mathrm{source}, \mathrm{target}\}$ represents a domain, $\lfloor\cdot\rfloor$ is the flooring function, and $\mathcal{H}(x)$ returns 1 when x > 0 and 0 otherwise. Here, $\{x_i^-\}_{i=1}^{N_-}$ and $\{x_j^+\}_{j=1}^{N_+}$ are normal and anomalous test clips in the domain d in the section n in the machine type m, respectively. N_- and N_+ are the numbers of normal and anomalous test clips in the domain d in the section n in the machine type m, respectively.

3.3. Experiment Results

Table 2: Detailed results for Fan.

Section (Domain)	AE		MobileNetV2		Our system	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
00 (source)	**66.69**	**57.08**	43.62	50.45	62.76	52.16
01 (source)	67.43	50.72	78.33	78.37	**85.68**	**81.95**
02 (source)	64.21	53.12	74.21	76.80	**75.61**	**77.42**
00 (target)	**69.70**	55.13	53.34	**56.01**	56.87	52.47
01 (target)	49.99	48.49	78.12	66.41	**92.83**	**89.84**
02 (target)	**66.19**	56.93	60.35	60.97	65.80	**74.26**
Arithmetic mean	64.03	53.58	64.66	64.84	**73.26**	**71.35**
Harmonic mean	63.24	53.38	61.56	63.02	**71.10**	**68.22**

Table 3: Detailed results for Gearbox.

Section (Domain)	AE		MobileNetV2		Our system	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
00 (source)	56.03	51.59	81.35	70.46	**84.21**	**71.51**
01 (source)	**72.77**	52.30	60.74	53.88	67.30	**59.02**
02 (source)	58.96	51.82	71.58	62.23	**82.12**	**66.93**
00 (target)	74.29	55.67	75.02	**64.77**	**77.87**	62.92
01 (target)	**72.12**	51.78	56.27	53.30	69.35	**60.28**
02 (target)	66.41	53.66	64.45	55.58	**79.15**	**72.49**
Arithmetic mean	66.76	52.80	68.24	60.03	**76.67**	**66.03**
Harmonic mean	65.97	52.76	66.70	59.16	**76.14**	**65.57**

Table 4: Detailed results for Pump.

Section (Domain)	AE		MobileNetV2		Our system	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
00 (source)	67.48	61.83	64.09	62.40	**76.53**	**62.74**
01 (source)	82.38	58.29	86.27	66.66	**94.61**	**86.47**
02 (source)	63.93	55.44	53.70	50.98	**71.59**	**56.63**
00 (target)	58.01	51.53	59.09	53.96	**59.64**	**55.95**
01 (target)	47.35	49.65	71.86	**62.69**	**72.86**	58.84
02 (target)	**62.78**	51.67	50.16	**51.69**	61.74	51.58
Arithmetic mean	63.66	54.74	64.20	58.06	**72.83**	**62.04**
Harmonic mean	61.92	54.41	61.89	57.37	**71.18**	**60.35**

In this part, performance of the proposed system is discussed and compared to the DCASE2021 task 2 baseline system. All results in the domain d in the section n in the machine type m are presented in Tables 2-8.

The AE baseline system models the normal features, so the abnormal features will have high reconstruction loss which is used as anomaly score. The MobileNetV2 baseline system train a classifier for 3 sections and the softmax value is used as the predicted probability for each section.

According to the AUC results shown in Table 2-8, the proposed system outperforms the baseline for all machines except ToyTrain. For the machine type of Fan, Gearbox, Pump, Slide rail, ToyCar and Valve, the arithmetic mean performance improved 8.6% (6.51%),

Table 5: Detailed results for Slide.

Section (Domain)	AE		MobileNetV2		Our system	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
00 (source)	74.09	52.45	61.51	53.97	**92.27**	**85.58**
01 (source)	82.16	60.29	79.97	55.62	**85.92**	**69.84**
02 (source)	78.34	65.16	**79.86**	**71.88**	77.70	66.98
00 (target)	**67.22**	**57.32**	51.96	51.96	67.18	57.11
01 (target)	**66.94**	53.08	46.83	52.02	62.06	**54.47**
02 (target)	46.20	50.10	55.61	**55.71**	**60.83**	53.09
Arithmetic mean	69.16	56.40	62.62	56.86	**74.33**	**64.51**
Harmonic mean	66.74	55.94	59.26	56.00	**72.48**	**62.74**

Table 6: Detailed results for ToyCar.

Section (Domain)	AE		MobileNetV2		Our system	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
00 (source)	**67.63**	51.87	66.56	**66.47**	65.95	58.37
01 (source)	61.97	51.82	71.58	**66.44**	**74.74**	66.16
02 (source)	**74.36**	**55.56**	40.37	47.48	61.36	47.58
00 (target)	54.50	50.52	61.32	**52.61**	**62.95**	52.32
01 (target)	64.12	52.14	72.48	63.99	**72.74**	**65.79**
02 (target)	56.57	52.61	45.17	48.85	**71.42**	**60.05**
Arithmetic mean	63.19	52.42	59.58	57.64	**68.19**	**58.38**
Harmonic mean	62.49	52.36	56.04	56.37	**67.82**	**57.56**

Table 7: Detailed results for ToyTrain.

Section (Domain)	AE		MobileNetV2		Our system	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
00 (source)	**72.67**	**69.38**	69.84	54.43	50.96	55.00
01 (source)	72.65	**62.52**	64.79	54.09	**72.86**	58.79
02 (source)	**69.91**	47.48	69.28	47.66	62.00	**48.74**
00 (target)	**56.07**	50.62	46.28	**51.27**	42.67	49.16
01 (target)	51.13	48.60	**53.38**	**49.60**	44.88	49.05
02 (target)	55.57	50.79	51.42	53.40	**57.60**	**53.68**
Arithmetic mean	**63.00**	**54.90**	59.16	51.74	55.16	52.40
Harmonic mean	**61.71**	**53.81**	57.46	51.61	53.31	52.14

Table 8: Detailed results for Valve.

Section (Domain)	AE		MobileNetV2		Our system	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
00 (source)	50.34	50.82	58.34	54.97	**62.42**	**59.53**
01 (source)	53.52	49.33	53.57	**50.09**	**61.30**	49.21
02 (source)	59.91	51.96	56.13	51.69	**77.47**	**62.53**
00 (target)	47.12	48.68	52.19	51.54	**58.97**	**62.89**
01 (target)	56.39	53.88	**68.59**	**57.83**	59.15	53.32
02 (target)	55.16	48.97	53.58	50.86	**69.56**	**52.68**
Arithmetic mean	53.74	50.61	57.07	52.83	**64.81**	**56.69**
Harmonic mean	53.41	50.54	56.51	52.64	**64.18**	**56.21**

8.43% (6%), 8.63% (3.98%), 5.17% (7.65%), 5% (0.74%), 7.74% (3.86%) in AUC (pAUC), respectively, compared with the best performing baseline. Our system can obtain signal that is more related to this task through denoising network and the overfitting model performs this task better than two baseline systems. However, for the machine type of ToyTrain, the arithmetic mean performance reduced 7.84% (2.5%) in AUC (pAUC) compared with the best performing baseline, mainly because of the model is not overfitting as other model by viewing the loss curve, due to the same setting of training hyperparameter for all types machines.

Note that the target domain performance is better than source domain performance in Fan01, Gearbox01 and ToyCar02, which may be caused by denoising. More experiment will be done to verify it in future work.

4. SUBMISSIONS

We generated our last submission by combining the proposed system and DCASE 2021 task 2 AE-baseline system. For machine tpye of Fan, Gearbox, Pump, Slide rail, ToyCar and Valve, the proposed system is used. For machine type of ToyTrain, the proposed system has a litter bit performance than baseline system, so we follow the baseline system but to train model for each section in machine type.

5. CONCLUSION

In this paper, a denoising-detection system is proposed to perform DCASE 2021 Task 2 by: (1) removing noise in each recording to obtain signal that is more related to this task; (2) training an overfitting model by leveraging the information of section. The proposed method significantly outperforms the baseline systems. In future work, we will consider (1) the effectiveness of different denoising methods and (2) the effectiveness of different classifiers to further enhance the performance of proposed system.

6. REFERENCES

[1] 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, 2018.

[2] T. Hayashi, T. Komatsu, R. Kondo, T. Toda, and K. Takeda, “Anomalous sound event detection based on wavenet,” in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2494–2498.

[3] D. W. Scott, “Outlier detection and clustering by partial mixture modeling,” in *COMPSTAT 2004—Proceedings in Computational Statistics*. Springer, 2004, pp. 453–464.

[4] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[6] A. Nicolson and K. K. Paliwal, “Deep learning for minimum mean-square error approaches to speech enhancement,” *Speech Communication*, vol. 111, pp. 44–55, 2019.

[7] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, “Deepmmse: A deep learning approach to mmse-based noise power spectral density estimation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1404–1415, 2020.

[8] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, “Anomalous sound detection based on interpolation deep neural network,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 271–275.

[9] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, “Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions,” *In arXiv e-prints: 2106.04492, 1–5*, 2021.

[10] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, and Y. Kawaguchi, “Mimii due: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions,” *arXiv preprint arXiv:2105.02702*, 2021.

[11] 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,” *arXiv preprint arXiv:2106.02369*, 2021.

[12] J. M. Lobo, A. Jiménez-Valverde, and R. Real, “Auc: a misleading measure of the performance of predictive distribution models,” *Global ecology and Biogeography*, vol. 17, no. 2, pp. 145–151, 2008.