ANOMALOUS SOUND DETECTION WITH THREE-SUBNETWORKS AND PRE-TRAINED MODELS

Technical Report

Ting Wu[1,2], *Jian Wen*[1,2], *Zhaoli Yan*[1,2,3*], *Xiaobin Cheng*[1,2]

[1] Key Laboratory of Noise and Vibration Research, Institute of Acoustics,
Chinese Academy of Sciences, Beijing, China
[2] School of Electronic, Electrical and Communication Engineering,
University of Chinese Academy of Sciences, Beijing, China
[3] College of Mechanical and Electrical, Beijing University of Chemical Technology, Beijing, China
{wuting, wenjian, zl_yan, xb_cheng}@mail.ioa.ac.cn

ABSTRACT

Unsupervised pretrained models have been used successfully in a wide range of scenarios. This report presents our work for DCASE 2024 Task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring. To solve this problem, a three-subnetworks is designed specifically for outlier exposure. The sample information is fully exploited to extract its embedding using classification networks as an auxiliary task, and then anomaly scores are calculated using clustering. Several pre-trained large models are fine-tuned with datasets from the DCASE 2024 challenge Task2 to further improve the performance. The ensemble of the above methods achieves an official score of 65.56% on the development dataset, being significantly superior to the baseline model's performance.

Index Terms— Anomalous sound detection, embedding extraction, pre-trained model

1. INTRODUCTION

In recent years, anomalous sound detection has become one of the key tasks in the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [1]. Since its inception in 2020, the task has progressively incorporated requirements relative to real-world scenarios. Firstly, models are trained exclusively using normal signals [2]. Secondly, anomalies must be detected despite the presence of domain shifts between the training and testing datasets [3]. Thirdly, the machine types in the development dataset are different from those in the evaluation dataset, preventing the fine-tuning of networks for specific target types. In addition, for each type of machine, only a single instance is provided [4]. Finally, the first-shot task is introduced this year where the attribute information of the machine acoustic signals is no longer fully available [5], [6], [7], [8].

In response to the challenges in anomalous sound detection, two primary methods have emerged. The first method is based on autoencoder (AE) models, known as inlier modeling (IM). This approach uses the reconstruction error of the AE model to identify anomalies. However, AE models often struggle to reconstruct anomalous signals effectively and face difficulties in handling signals with domain shifts. Due to the complex acoustic background, AE-based methods often fail to capture the subtle features of signal anomalies [9]. Consequently, a popular approach based on discriminative models, known as outlier exposure (OE), has been widely used recently. This method utilizes a classification network as an auxiliary task to extract signal embedding vectors and further use clustering or other techniques to detect anomalies. By applying deep networks to capture detailed signal information, this approach has achieved outstanding results in recent years [10], [11], [12].

Based on the problems mentioned above, this paper proposes an anomalous sound detection method that incorporates three sub-networks and a pre-trained model. Our approach is designed to extract and process date from the time domain, frequency domain, and time-frequency domain. The network architecture is redesigned based on the work of [13]. Previous studies have shown that fully utilizing information such as device ID, speed microphone position, and operating environment can significantly enhance the extraction of signal representations [14]. After a thorough evaluation of Gaussian Mixture Models (GMM) and K-Means, we decided on K-Means as the anomaly score calculator finally. Considering the advancements of pre-trained models, we further trained a pre-trained model and finally integrated its results with the outputs of the three sub-networks.

The rest of the paper is organized as follows. Section 2 describes the three sub-networks. Section 3 covers the pre-trained models. Section 4 presents the ensemble strategy and the results. Finally, Section 5 concludes the work.

2. THREE-SUBNETWORKS

In this section, we propose a detection method using a multi-channel classification network called three sub-networks. Acoustic samples are classified according to their attributes and categories, and a network is trained for classification task. The embedding vectors are then extracted as low-dimensional representations of signals to calculate the anomaly score. The details of the network are as follows.

During the front-end processing, the length of all signals is adjusted to 12 seconds to ensure consistent data length. Subsequently, three features are extracted from the signals: the raw waveform, the amplitude spectrum, and the Short-Time Fourier Transform (STFT) spectrogram. The window length and frame shift of STFT are set to 1024 and 512, respectively. Data augmentation techniques such as mixup are applied by linearly interpolating. The Adam optimizer is employed with a learning rate of 1e-3 and a batch size of 64.

In terms of the network model, the subnetwork for the raw waveform and amplitude spectrum features consists of three one-dimensional convolutional layers followed by a Flatten layer and four fully connected layers, using batch normalization for feature extraction. Both subnetworks ultimately output 128-dimensional embeddings. For the STFT spectrogram features, the subnetwork is a modified ResNet composed of four basic blocks, followed by a Flatten layer, with batch normalization applied to standardize the data distribution. This subnetwork finally outputs a 256-dimensional embedding.

The output vectors from the three subnetworks are concatenated to form a 512-dimensional embedding. The loss function chosen is the sub-cluster AdaCos loss, which uses a dynamic adaptive scaling parameter and multiple class centers. Previous studies [15] have demonstrated the superior performance of this loss function in anomalous sound detection tasks. The network model is shown in Figure 1:

3. PRE-TRAINED MODELS

In this section, five pre-trained models including HuBERT, BEATs, Wav2Vec 2.0, WavLM, and UniSpeech-SAT are introduced briefly.

HuBERT is a pre-trained self-supervised learning (SSL) model designed to learn audio representations from unlabeled raw signals to perform various audio tasks [16]. HuBERT uses an offline clustering algorithm to generate pseudo-labels and employs a stacked CNNs model to extract features from the raw signals. Finally, the audio input is predicted through transformer layers. The training of HuBERT involves two steps: in the first step, pseudo-labels are generated using Mel-frequency cepstral coefficients (MFCC); in the second step, labels are generated by using the embeddings produced in the first step.

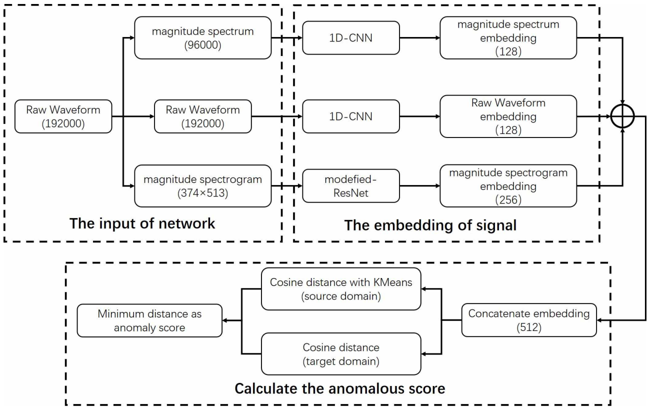
Raw Waveform (192000)
magnitude spectrum (96000)
Raw Waveform (192000)
magnitude spectrogram (374×513)
The input of network
1D-CNN
1D-CNN
modefied-ResNet
magnitude spectrum embedding (128)
Raw Waveform embedding (128)
magnitude spectrogram embedding (256)
The embedding of signal
Concatenate embedding (512)
Cosine distance with KMeans (source domain)
Cosine distance (target domain)
Minimum distance as anomaly score
Calculate the anomalous score

Figure 1: The three-SubNetworks framework.

BEATs is an iterative audio pre-trained SSL model. The model is trained using random projections as audio labels and enhanced by pretraining or fine-tuning [17]. In each iteration, discrete labels of unlabeled audio are generated using an audio tagger. These labels are used to optimize the SSL model with masking and discrete labels. After convergence, the SSL model serves as a teacher model to guide the acoustic representation and learn audio semantics through knowledge distillation.

Wav2Vec 2.0 is an SSL speech representation learning framework proposed by Facebook [18]. It consists of a multi-layer convolutional feature encoder, a quantization module, and several transformer layers. The feature encoder transforms raw audio into latent representations of the audio. The quantization module further discretizes these latent representations, which are considered as acoustic targets. Meanwhile, the latent representations are randomly masked and input into the transformer layers.

WavLM is a pretraining framework that utilizes masked speech denoising and prediction [19]. Some of the inputs are overlap masks or simulated noise, with the objective of predicting pseudo-labels of the original audio in the masked regions. The WavLM model comprises a convolutional encoder and a transformer. During training, WavLM randomly transforms the input waveforms, masks 50% of the audio signal, and predicts the labels corresponding to the masked positions at the output.

UniSpeech-SAT is a speaker-aware pretraining model based on UniSpeech [19]. This model uses a multi-task learning framework with a contrastive loss, integrating sentence-level contrastive loss with SSL objectives. Additionally, the model uses a sentence-level mix strategy for speech enhancement.

4. RESULT AND ENSEMBLE STRATEGY

Table 1 presents the results of the five pre-trained models. Table 2 compares the baseline results with those of the proposed models using GMM and K-Means anomaly score calculators.

Table 1: Pre-training model results.

	Method	HuBERT /%	BERTs /%	Wav2Vec 2.0/%	WavLM /%	UniSpeech -SAT/%
Bearing	AUC(source)	64.2	**68.39**	61.67	64.64	60.64
	AUC(target)	**72.48**	63.44	65.32	71.4	65.04
	pAUC	58.47	56.36	60.31	**60.73**	56.36
Fan	AUC(source)	55.6	51.15	55.92	55.43	**56.75**
	AUC(target)	63.84	**73.76**	51.56	70.6	62.55
	pAUC	51.89	56.63	54.47	**60.57**	52.47
Gearbox	AUC(source)	80.24	**85.83**	65.56	47.43	53.4
	AUC(target)	80.24	**84.2**	69.2	52.76	67.11
	pAUC	61.78	**70.57**	57.73	50.94	53.94
Slider	AUC(source)	61.08	**64.96**	50.52	55.6	53.64
	AUC(target)	54.96	**58.36**	58.24	51.4	56.6
	pAUC	48.73	**51.89**	48.63	49.89	50.94
ToyCar	AUC(source)	50.72	44.76	53	**58.11**	46.32
	AUC(target)	47.31	**56.4**	43.99	40.08	45.99
	pAUC	47.94	48.94	**51.10**	48.89	48.78
ToyTrain	AUC(source)	74.12	**84.2**	67.28	80.51	83.71
	AUC(target)	**62.68**	41.2	51.4	52.28	45.52
	pAUC	51.84	48.36	49.89	**53.21**	51
Valve	AUC(source)	86.76	83.04	87.2	77.27	**87.68**
	AUC(target)	**69.88**	43.31	59.52	66.88	55.04
	pAUC	63.47	55.63	**70.95**	57.10	64.52
All	AUC(source)	**65.34**	65.13	61.21	60.79	60.12
	AUC(target)	**62.77**	56.71	55.86	55.72	55.64
	pAUC	54.27	**54.72**	54.48	54.09	53.61
hmean	score	**60.41**	58.52	57.04	56.73	56.33

The ensemble strategy employed in this paper is a mean aggregation method. Given that the varying scores obtained from different models, they are normalized firstly. The final anomaly score is computed through ensemble aggregation based on specified proportions.

In this paper, four subsystems are proposed as follows. Since the development dataset and the evaluation dataset are significantly different in terms of machine types, each ensemble is designed accordingly. Ensemble-1 combines the proposed three-subnetworks. Ensemble-2 combines the proposed network with results from the five pre-trained models. Ensemble-3 integrates the proposed network model with the best-performing HuBERT model among the five pre-trained models. Ensemble-4 filters out models with low accuracy before combining the networks. Table 3 shows the results of these four ensembles proposed in this paper.

Table 2: Baseline and Three-SubNetwork model results.

	Method	Baseline MSE/%	Baseline MAHALA/%	Proposed GMM/%	Proposed KMeans/%
Bearing	AUC(source)	62.01	54.43	61.72	**65.8**
	AUC(target)	61.4	51.58	69.99	**73.31**
	pAUC	57.58	57.58	**59.57**	62
Fan	AUC(source)	67.71	**79.37**	63.44	58.04
	AUC(target)	55.24	42.7	**68**	63.8
	pAUC	**57.53**	53.44	56.31	56.89
Gearbox	AUC(source)	70.4	**81.82**	71.16	70.8
	AUC(target)	69.34	**74.35**	74.4	71.72
	pAUC	55.65	55.74	**57.78**	55.94
Slider	AUC(source)	66.51	75.35	97.52	**98.8**
	AUC(target)	56.01	68.11	93.04	**95.84**
	pAUC	51.77	49.05	75.05	**89.31**
ToyCar	AUC(source)	**66.98**	63.01	52.96	55.51
	AUC(target)	33.75	37.35	45.12	**54.6**
	pAUC	48.77	**51.04**	49.52	48.42
ToyTrain	AUC(source)	**76.63**	61.99	63.48	48.87
	AUC(target)	46.92	39.99	60.36	**62.04**
	pAUC	47.95	48.21	53	**53.05**
Valve	AUC(source)	51.07	55.69	95.31	**95.64**
	AUC(target)	46.25	53.61	51.88	**68.68**
	pAUC	52.42	51.26	62.78	**70.05**
All	AUC(source)	64.95	65.77	**69.03**	66.28
	AUC(target)	50.27	49.59	62.94	**67.91**
	pAUC	52.84	52.28	58.25	**60.05**
hmean	score	55.33	55.04	63.10	**64.57**

Table 3: The result of ensemble.

	Method	Ensemble -1/%	Ensemble -2/%	Ensemble -3/%	Ensemble -4/%
Bearing	AUC(source)	65.68	**67.32**	66.08	67.03
	AUC(target)	71.84	73.96	73.6	**74.52**
	pAUC	59.78	59.89	59.47	**60.26**
Fan	AUC(source)	**61.76**	57.36	58.6	59.76
	AUC(target)	66.4	67.68	66.63	**69.64**
	pAUC	55.52	**56.21**	55.42	56.15
Gearbox	AUC(source)	71.08	76.44	75.87	**81.36**
	AUC(target)	74.32	79.4	79.24	**82.08**
	pAUC	57.68	60.05	60	**63.68**
Slider	AUC(source)	**96.72**	94.91	95.2	94.56
	AUC(target)	**94.72**	91.16	92.39	87.96
	pAUC	**80.47**	66.15	73.21	66.94
ToyCar	AUC(source)	48.68	49.52	48.84	**49.64**
	AUC(target)	50.56	50.16	50.08	**52.27**
	pAUC	**49.21**	48.31	48.63	48.26
ToyTrain	AUC(source)	60.96	69.52	66.24	**72.08**
	AUC(target)	60.68	59.36	61.92	**63.28**
	pAUC	53.26	53.15	53.84	**53.84**
Valve	AUC(source)	95.4	95.72	95.16	**95.88**
	AUC(target)	55.32	55.76	56.68	**64.12**
	pAUC	64.21	64.36	64.36	**66**
All	AUC(source)	67.79	69.35	68.65	**70.73**
	AUC(target)	65.22	65.68	66.16	**68.72**
	pAUC	**58.74**	57.7	58.4	58.58
hmean	score	63.68	63.86	64.1	**65.56**

5. CONCLUSIONS

In this paper, we propose a three-subnetwork structure for anomalous sound detection. Classification networks, being trained from multiple signal perspectives, are used as auxiliary tasks to extract signal embeddings. Additionally, data augmentation technique is included and several large pre-trained models are integrated. Compared to the AE Baseline, this method shows significant improvements in both AUC and pAUC metrics. Four results of distinct ensemble approach are given.

6. REFERENCES

[1] J. Yan, Y. Cheng, Q. Wang, L. Liu, W. Zhang, and B. Jin, "Transformer and graph convolution-based unsupervised detection of machine anomalous sound under domain shifts," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[2] 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," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219559355

[3] 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," *ArXiv*, vol. abs/2106.04492, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235367775

[4] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on dcase 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *ArXiv*, vol. abs/2305.07828, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258685340

[5] 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.

[6] 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.

[7] 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 *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[8] 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.

[9] K. Wilkinghoff and F. Kurth, "Why do angular margin losses work well for semi-supervised anomalous sound detection?" *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[10] Y. Zeng, H. Liu, L. Xu, Y. Zhou, and L. Gan, "Robust anomaly sound detection framework for machine condition monitoring," *Proc. DCASE*, 2022.

[11] K. Wilkinghoff, "Fraunhofer fkie submission for task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," Technical report, DCASE2023 Challenge, Tech. Rep., 2023.

[12] S. Chen, J. Wang, J. Wang, and Z. Xu, "Mdam: Multi-dimensional attention module for anomalous sound detection," in *International Conference on Neural Information Processing*. Springer, 2023, pp. 48–60.

[13] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 276–280.

[14] J. Guan, F. Xiao, Y. Liu, Q. Zhu, and W. Wang, "Anomalous sound detection using audio representation with machine id based contrastive learning pretraining," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] C. Yu, X. Su, and Z. Qian, "Multi-stage audio-visual fusion for dysarthric speech recognition with pre-trained models," *IEEE Transactions on Neural Systems and*

Rehabilitation Engineering, vol. 31, pp. 1912–1921, 2023.

[16] X. Yang, F. Lv, F. Liu, and G. Lin, “Self-training vision language berts with a unified conditional model,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[17] R. Fukuda, K. Sudoh, and S. Nakamura, “Improving speech translation accuracy and time efficiency with fine-tuned wav2vec 2.0-based speech segmentation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[18] H. Huang, L. Wang, J. Yang, Y. Hu, and L. He, “W2vc: Wavlm representation based one-shot voice conversion with gradient reversal distillation and ctc supervision,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 45, 2023.

[19] A. L. Zorrilla, M. I. Torres, and H. Cuayáhuitl, “Audio embedding-aware dialogue policy learning,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 525–538, 2022.