THUEE SYSTEM FOR FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION FOR MACHINE CONDITION MONITORING

Technical Report

Anbai Jiang[1]*, Qijun Hou*[1]*, Jia Liu*[1]*, Pingyi Fan*[1]*, Jitao Ma*[2]*,*
Cheng Lu[2]*, Yuanzhi Zhai*[1]*, Yufeng Deng*[1]*, Wei-Qiang Zhang*[1]

[1] Department of Electronic Engineering, Tsinghua University, Beijing, China
[2] School of Economics and Management, North China Electric Power University, Beijing, China
{jab22, hqj19}@mails.tsinghua.edu.cn, {liuj, fpy}@tsinghua.edu.cn,
{majitao_w, lucheng1983}@163.com, {diyz19, dyf20}@mails.tsinghua.edu.cn, wqzhang@tsinghua.edu.cn

ABSTRACT

This report presents our work for DCASE 2023 Task 2: first-shot unsupervised anomalous sound detection for machine condition monitoring. This task mainly focuses on first-shot problems compared with previous challenges. No hyperparameter tuning and developing systems on some machines while testing on other machines bring a lot of challenges. We have developed several kinds of systems to detect first-shot sound anomalies better: training embedding extraction systems from scratch, finetuning pre-trained embedding extractors, and employing normalizing flows. Different kinds of systems give complementary information. We achieve the best hmean of 69.46% on the development set through system fusion.

Index Terms— Anomaly detection, sound, embedding extraction, normalizing flows

1. INTRODUCTION

Recent years saw a dramatic improvement in Artificial Intelligence (AI) technologies and the accelerating Internet of Things (IoT) deployment. The concept of the Artificial Internet of Things (AIoT), a combination of AI and IoT, has become a heated research topic. The ubiquitous sensors and devices form hierarchical networks, generating numerous data continuously on which powerful AI technologies can be deployed. This scheme can realize more and more applications. Industrial manufacturing is one of the most probable fields that benefit from AIoT since numerous sensors can be deployed in the production site, continuously monitoring the working status, while anomaly detection techniques can be leveraged to identify potential machine failures, leading to a great improvement in efficiency and safety. However, challenges remain for anomaly detection in industrial scenarios, especially audio-based anomaly detection. We believe the main challenges can be summarized as follows:

1. Lack of anomalous samples for training. Anomalies are rare to happen, and limited anomalies are utilized as validation. Anomaly detectors must be trained without anomalous samples, and models must be either unsupervised or trained by a proxy task. This lack of direct supervision brings huge challenges for anomaly detection.
2. The noise mixed with the machine audio. Recorded audio of a specific machine is often mixed with various kinds of background noise: the sound of other machines, the sound of human activities, etc. The background noise is also variable over time and likely identified as an anomaly.
3. Domain shift caused by the variational working conditions. Working conditions continuously change, and different working conditions correspond to different patterns, requiring the model to be generalized in all scenarios. To feature this issue, domain shift is introduced in the challenge [1, 2, 3], where most normal clips for training are from the source domain.

In this paper, we describe the THUEE system for first-shot unsupervised anomalous sound detection for machine condition monitoring. Multiple classification models and a probabilistic model are developed for the challenge, and all four submitted systems are combinations of these models. We will introduce these models individually, then present the composition of four ensemble systems.

2. MOBILEFACENET-BASED CLASSIFIER

In this subsection, we provide a classification model, MobileAnoNet(MAN), a network trained by supervised classification. MAN combines a front-end feature extractor implemented by a modified MobileFaceNet(MFN) [4] and a back-end KNN [5] anomaly detector.

MobileAnoNet is supervised by both the machine type and working condition labels simultaneously. Multiple parallel classification heads corresponding to machine type and working condition labels are employed to train the feature extractor. Each classification head consists of an independent full-connection layer and a Cross-Entropy loss. The overall loss function is defined as Eq.(1), where $\mathcal{L}_{machine}$ and $\mathcal{L}_{condition}$ are the loss given by the machine type label and the working condition labels.

$$\mathcal{L} = \mathcal{L}_{machine} + \mathcal{L}_{condition} \tag{1}$$

$\mathcal{L}_{machine}$ is implemented by Center Loss [6], formulated as Eq.(2), where the first term is the cross entropy loss between each predicted machine type $W_m^T x_i$ and the ground truth y_i^m. W_m^T is the weight of the classification head. The second term is the Mean Square Loss (MSE) between each embedding x_i and the center of the corresponding class $c_{y_i^m}$, where λ is a hyper-parameter to balance two terms.

Machine	AUCs	AUCt	pAUC	hmean
bearing	65.6	53.1	52.9	56.6
fan	86.2	61.6	63.7	68.9
gearbox	87.0	79.3	70.3	78.3
slider	97.1	96.3	85.8	92.8
ToyCar	67.3	40.0	49.1	49.8
ToyTrain	52.5	42.3	48.8	47.5
valve	76.2	66.5	51.8	63.2
all_hmean	73.2	57.5	58.0	62.1

Table 1: Performance of MobileAnoNet(MAN)

Machine	AUCs	AUCt	pAUC	hmean
bearing	67.66	64.64	51.84	60.56
fan	92.40	84.00	76.32	83.73
gearbox	77.74	75.02	57.68	68.92
slider	91.00	85.54	66.37	79.48
ToyCar	72.36	44.44	50.37	53.40
ToyTrain	52.38	47.16	48.32	49.19
valve	69.20	66.60	53.58	62.33
hmean	72.30	63.00	56.45	63.26

Table 2: Performance of WSP-NFCDEE

$$\begin{aligned}\mathcal{L}_{machine} &= \mathcal{L}_s + \lambda\mathcal{L}_c \\ &= \frac{1}{n}\sum_{i=1}^{n} CE(W_m^T x_i, y_i^m) + \lambda \cdot \frac{1}{n}\sum_{i=1}^{n} \|x_i - c_{y_i^m}\|_2^2\end{aligned} \tag{2}$$

$\mathcal{L}_{condition}$ is the sum of the cross entropy loss of working conditions, which is formulated as Eq.(3). k_i is the number of working condition labels x_i possesses, c_j is the j th working conditions of x_i and $W_{c_j}^T x_i$ is the predicted working condition of c_j.

$$\mathcal{L}_{condition} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k_i} CE(W_{c_j}^T x_i, y_i^{c_j}) \tag{3}$$

The Short-Time Fourier Transform(STFT) spectrogram is selected as the input feature. An embedding vector is calculated for each audio clip by averaging the embedding vector of every frame by the trained feature extractor. A KNN model is trained by all the embedding vectors from the train set.

We train our MAN on all the training data from the development and Additional training dataset. The performance(AUCs, AUCt, pAUC) is shown in Table 1.

3. NF-CDEE

In this subsection, we provide a detailed description of our probabilistic model: WSP-NFCDEE. WSP-NFCDEE is an extension of the NFCDEE model [7], incorporating a Weighted Statistic Pooling (WSP) layer before the normalizing flows. This modification significantly enhances performance across various machine types, particularly on the slider. Consequently, we refer to this modified approach as WSP-NFCDEE.

Let $x \in \mathbb{R}^{M\times T}$ denote the mel-spectrogram, where M represents the number of Mel bins, and T denotes the number of frames. The WSP module computes the mean vector $y \in \mathbb{R}^M$ and the standard deviation vector $z \in \mathbb{R}^M$ of X along the time axis. The output of the WSP module is obtained by combining $\alpha \cdot y$ and $\beta \cdot z$, where α and β are two trainable parameters that conform to the constraints:

$$\alpha + \beta = 1, \alpha, \beta > 0 \tag{4}$$

This integration of the WSP layer preceding the normalizing flows improves the overall performance, making WSP-NFCDEE a powerful choice for a wide range of machine types, with particular emphasis on the slider.

The log mel-spectrogram is selected as the input feature for both WSP-NFCDEE and IMDN models. Specifically, the input feature is obtained through the Short-Time Fourier Transform (STFT) with Mel scaling, which effectively captures the non-linear frequency characteristics of the audio signal.

While WSP-NFCDEE performs well on a single type of machine due to limited data, the system trained on a single machine may not have sufficient generalization ability on new data. Therefore, the machines are partitioned into four groups based on their signal features:

1. Stationary: bansaw, shaker, bearing, fan, ToyCar
2. Non-stationary: ToyDrone, ToyNscale, ToyTank, ToyTrain, Vacuum
3. Periodic: gearbox, slider
4. Aperiodic with impulse: grinder, valve

Four different WSP-NFCDEE models are trained on each group to enhance the generalization ability of the trained models. The performance on the development set is presented in Table 2. Results from the grouping method outperform the performance of training all machines together. In particular, for certain machine types, such as fans, the performance improvement of the grouping method over a single machine is significant.

4. PRE-TRAINED AUDIO MODELS

As known, Pre-trained Language Models (PLMs) have demonstrated powerful capability and great potential. We also investigate the use of pre-trained audio models in this year's challenge.

4.1. Pre-trained Models

Four types of pre-trained models are employed in the scheme: Wav2Vec 2.0 [8], HuBERT [9], Unispeech [10], and WavLM [11], most of which are pre-trained on speech datasets. Wav2Vec 2.0 improves Wav2Vec [12] by employing a transformer encoder as the context network. HuBERT adopts the architecture of Wav2Vec 2.0 while utilizing the masked language modeling task proposed in BERT [13], as well as introducing a novel clustering algorithm to process the mel-spectrogram. Unispeech incorporates multi-task learning and improves the performance in multi-lingual speech recognition and domain transfer tasks in audio. WavLM combines HuBERT with multiple kinds of data augmentation, effectively promoting the general performance on multiple speech-related downstream tasks. All models are implemented by PyTorch [14] and HuggingFace. We use the XLS-R 300M version for Wav2Vec, HuBERT-large, Unispeech-large, and WavLM-large, each of which contains approximately 300M parameters.

Table 3: Performance of Pre-trained Models

	Wav2Vec			HuBERT			Unispeech			WavLM		
	mean-min	tf-none	tf-min	mean-min	tf-none	tf-min	mean-min	tf-none	tf-min	mean-min	tf-none	tf-min
bearing	62.62	62.52	64.02	71.29	66.89	71.19	74.74	73.62	74.90	71.70	65.50	71.14
fan	66.66	60.53	64.77	59.45	61.41	62.29	56.92	49.71	57.39	55.70	48.25	55.87
gearbox	77.77	68.9	71.31	67.61	65.22	69.70	73.49	70.59	69.67	74.65	77.14	75.89
slider	83.96	87.63	83.92	80.82	78.63	77.99	80.87	84.22	85.02	82.88	80.38	86.42
ToyCar	58.92	59.90	60.01	56.09	62.00	59.52	57.26	56.46	57.36	55.20	57.86	57.53
ToyTrain	55.92	56.94	56.53	54.79	53.80	53.02	54.40	56.17	56.96	61.43	55.36	58.75
valve	68.61	59.57	67.08	61.33	56.15	58.71	67.64	60.16	65.93	66.07	50.94	58.19
hmean	66.56	63.94	65.88	63.41	62.61	63.66	65.08	62.57	65.39	65.49	60.15	64.64

The harmonic mean of AUCs, AUCt, and pAUC of each machine type is presented in this table, where mean and tf denote the mean pooling and the transformer aggregation, respectively, and none and min denote the regular detection and the detection with soft scoring.

Table 4: Performance of Ensemble Models

	Ensemble-1				Ensemble-2				Ensemble-3				Ensemble-4			
	AUCs	AUCt	pAUC	hmean	AUCs	AUCt	pAUC	hmean	AUCs	AUCt	pAUC	hmean	AUCs	AUCt	pAUC	hmean
bearing	67.66	64.64	51.84	60.56	78.80	65.04	56.00	65.33	81.00	69.42	56.53	67.50	78.70	67.86	54.74	65.63
fan	92.40	84.00	76.32	83.73	98.24	64.32	65.37	73.13	94.94	78.98	62.89	76.74	98.64	77.58	63.58	77.41
gearbox	77.74	75.02	57.68	68.92	85.16	83.58	68.16	78.17	86.88	85.70	68.79	79.54	90.60	88.68	73.89	83.69
slider	91.00	85.54	66.37	79.48	99.90	93.62	84.21	92.12	99.52	94.14	83.74	91.99	99.34	93.22	81.68	90.81
ToyCar	72.36	44.44	50.37	53.40	66.52	56.72	49.53	56.76	63.14	63.34	48.63	57.48	62.14	59.24	48.42	55.94
ToyTrain	52.38	47.16	48.32	49.19	69.84	51.20	49.53	55.51	67.72	57.60	48.53	56.89	63.96	60.38	50.58	57.73
valve	69.20	66.60	53.58	62.33	79.28	62.70	52.74	63.13	75.98	69.92	54.47	65.47	78.96	78.16	55.79	69.15
hmean	72.30	63.00	56.45	63.26	80.84	65.62	58.80	67.24	79.36	72.31	58.53	68.94	79.24	73.02	59.26	69.46

All pre-trained models are fine-tuned with all the parameters on the development dataset and the additional training dataset, by classifying the attribute information. Each unique combination of the provided attributes is considered a unique category, resulting in 167 classes for the datasets. The input for each model is 2s segments randomly sliced from the 10s clip, and it is processed by the pre-trained models. Since the output shape of these models is too big for anomaly detection, statistical pooling modules are attached to the output of these models, which maps the output to a fixed size of 128. Additional linear classification heads then map the 128-dimension embedding to the logits. All models are trained by ArcFace [15]. We adopt AdamW as the optimizer with a learning rate of 5e-4. Models are trained for 10k steps and validated periodically. The best-performing models are saved for inference.

4.2. Anomaly Detection

After fine-tuning the models, we extract the 128-dimension embedding as the feature representation and conduct anomaly detection on these embeddings. However, since each 10s clip corresponds to multiple 2s segments, we investigate how to aggregate the segment embedding series into a clip embedding. For efficient representation learning of audios, we feed 2-second segments into pre-trained models and obtain at least five different embeddings with different window shifts for 10-second audio input. Then a pooling method is applied over different embeddings of a clip of audio. We investigate average pooling and adopt a 1-layer transformer to aggregate the features. Both methods yield a 128-dimension embedding as the clip embedding.

Anomaly detection is conducted on the 128-dimension clip embedding. We adopt KNN as the anomaly detector since it is much more robust across all machines than other detectors. The distance metric is chosen as cosine distance, and the number of neighbors k is selected as 2. KNN is trained on all the embeddings of the training set. However, the soft scoring algorithm introduced in the baseline systems [16] is also investigated for KNN, in which we train two KNN detectors by all the source embeddings and all the target embeddings, respectively. A query embedding is processed by two detectors, and the minimum score is selected as the anomaly score.

The performance of each single model is presented in Table 3.

5. ENSEMBLE

We ensemble all the proposed models by score-level. The performance of all four ensemble models is presented in Table 4. Ensemble-1 comprises MobileAnoNet and WSP-NFCDEE. Ensemble-2 comprises MobileAnoNet, WSP-NFCDEE, and all pre-trained models of transformer aggregation. Ensemble-3 comprises MobileAnoNet, WSP-NFCDEE, and all pre-trained models of transformer aggregation and soft scoring. Ensemble-4 comprises MobileAnoNet, WSP-NFCDEE, and all pre-trained models of mean pooling and soft scoring.

6. REFERENCES

[1] 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," *In arXiv e-prints: 2305.07828*, 2023.

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

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

[4] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 428–438.

[5] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 427–438.

[6] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer, 2016, pp. 499–515.

[7] J. A. Lopez, G. Stemmer, P. Lopez-Meyer, P. Singh, J. A. del Hoyo Ontiveros, and H. A. Cordourier, "Ensemble of complementary anomaly detectors under domain shifted conditions." in *DCASE*, 2021, pp. 11–15.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[10] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang, "Unispeech: Unified speech representation learning with labeled and unlabeled data," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 937–10 947.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[12] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[16] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," *In arXiv e-prints: 2303.00455*, 2023.