ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447183

EXPLORING LARGE SCALE PRE-TRAINED MODELS FOR ROBUST MACHINE ANOMALOUS SOUND DETECTION

Bing Han[1], Zhiqiang Lv[2*], Anbai Jiang[3], Wen Huang[1], Zhengyang Chen[1], Yufeng Deng[2],*
Jiawei Ding[2], Cheng Lu[4], Wei-Qiang Zhang[3], Pingyi Fan[3], Jia Liu[3], Yanmin Qian[1,†]

[1]Auditory Cognition and Computational Acoustics Lab
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Huakong AI Plus Company Limited, Beijing, China
[3]Department of Electronic Engineering, Tsinghua University, Beijing, China
[4]School of Economics and Management, North China Electric Power University, Beijing, China

ABSTRACT

Machine anomalous sound detection is a useful technique for various applications, but it often suffers from poor generalization due to the challenges of data collection and complex acoustic environment. To address this issue, we propose a robust machine anomalous sound detection model that leverages self-supervised pre-trained models on large-scale speech data. Specifically, we assign different weights to the features from different layers of the pre-trained model and then use the working condition as the label for self-supervised classification fine-tuning. Moreover, we introduce a data augmentation method that simulates different operating states of the machine to enrich the dataset. Furthermore, we devise a transformer pooling method that fuses the features of different segments. Experiments on the DCASE2023 dataset show that our proposed method outperforms the commonly used reconstruction-based autoencoder and classification-based convolutional network by a large margin, demonstrating the effectiveness of large-scale pre-training for enhancing the generalization and robustness of machine anomalous sound detection. In Task2 of DCASE2023, we achieve 2nd place with these methods.

Index Terms— machine anomalous sound detection, self-supervised pre-train, fine-tune

1. INTRODUCTION

The task of anomalous sound detection (ASD) [1] involves determining whether the sound produced by a specific machine should be classified as normal or anomalous. It has garnered significant attention from researchers owing to its critical role in guaranteeing the safety and operational effectiveness of industrial machinery.

Unlike acoustic scene classification, this task is an unsupervised learning scenario where training data comprises only normal-state samples. The objective is to determine whether a given test sample belongs to another class known as the anomaly class, which encompasses various anomalous situations.

Recently, there has been a notable emergence of deep learning methodologies designed for ASD. And these methods can be categorized into two primary groups based on their training criteria: unsupervised and self-supervised techniques. Unsupervised approaches [2, 3, 4] aim to bolster ASD model efficiency by estimating the distribution of normal sounds and subsequently detecting anomalies. This detection process hinges on evaluating the likelihood of unknown sounds conforming to this established normal distribution.

On the other hand, self-supervised techniques [5, 6, 7] leverage the metadata of audio files (such as machine types and machine status) as labels to train a classifier to extract the latent representation of machine sounds. Despite the notable achievements of these methodologies, their capacity for generalization often remains limited, primarily due to the inherent challenges posed by data collection and the complexity of acoustic environments.

To address this challenge, our focus is on identifying robust audio encoders with the capacity for generalization, thereby mitigating the risk of overfitting to a limited training dataset. In recent years, the adoption of large-scale pre-trained models has emerged as the predominant approach for achieving state-of-the-art performance. Building upon the groundbreaking achievements of models like BERT [8] , researchers within the speech community have introduced a range of innovative architectures. Examples include wav2vec 2.0 [9], HuBERT [10], UniSpeech [11] and WavLM [12], all of which leverage extensive unlabeled speech data. These methods have yielded impressive results in automatic speech

*Equal Contribution
†Yanmin Qian is the corresponding author

This work was supported in part by China NSFC projects under Grants 62122050, 62071288 and 62276153, in part by Shanghai Municipal Science and Technology Commission Project under Grant 2021SHZDZX0102.

recognition (ASR) tasks, capitalizing on the power of pre-training and demonstrating the potential of leveraging vast amounts of unlabeled data in the speech domain.

Inspired by the excellent results of the pre-trained models in various downstream tasks [13, 14], we adopt multiple large scale pre-trained models to enhance the generalization performance in the context of anomalous sound detection. To address the issue of limited data diversity from a single machine, we introduce a novel approach known as "status augmentation," which simulates diverse machine statuses. Finally, we enhance the system's stability by incorporating a transformer pooling method for segment fusion. We participated in Task2 of DCASE and achieved 2nd place with these strategies. Our contribution can be succinctly summarized as follows::

- We **firstly** explore several large scale pre-trained models for robust machine anomalous sound detection.
- We are the **first** to propose a data augmentation method named "status augmentation" for anomalous sound detection to simulate different operation status of machine by perturbing the speed.
- We adopt transformer pooling to gather the embeddings from the same recording into one embedding for discovering effective information.

2. LARGE SCALE PRE-TRAINED MODELS

In this section, we will provide a concise overview of several large scale pre-trained models investigated in this paper.

Wav2Vec 2.0 [9] is a continuation of the wav2vec series. It replaces the original architecture's convolutional context network with multi-layer transformer-based encoder. While incorporating discrete speech units and a quantization module akin to the vq-wav2vec model [15], wav2vec 2.0 reverts to the original contrastive objective used in the first version of wav2vec, rather than adopting BERT's masked language modeling objective. It's worth noting that we use the scale-up XLS-R version, which utilize much more training data.

UniSpeech [11] presents a multi-task model that integrates a self-supervised learning objective, similar to wav2vec 2.0, with a supervised ASR objective using Connectionist temporal classification. This combined approach enables enhanced alignment between discrete speech units and the phonetic structure of the audio, resulting in improved performance in multi-lingual speech recognition and audio domain transfer tasks.

HuBERT [10] utilizes the architecture of wav2vec 2.0 while substituting the contrastive objective with BERT's original masked language modeling objective. This transformation involves a two-step pre-training process. In the clustering step, short speech segments are assigned pseudo-labels, and in the prediction step, the model is trained to predict these pseudo-labels at randomly-masked positions within the original audio sequence. This strategy facilitates the incorporation of BERT's objective into the wav2vec 2.0 architecture.

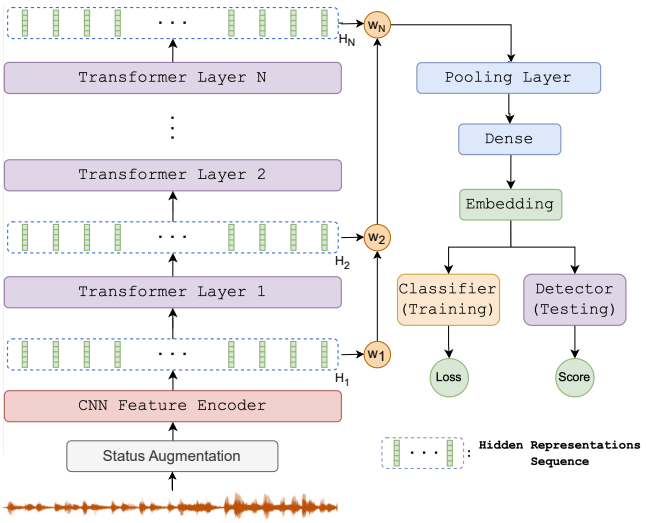

Fig. 1. Overview of the pretrained models based classification system. For the waveform sound, we first augment it with status augmentation. Then, the augmented waveform is fed into a large-scale pre-trained model and transformed into embedding for training and evaluation.

WavLM [12] models adhere to the HuBERT framework but emphasize data augmentation during the pre-training phase to enhance speaker representation learning and subsequently improve performance in speaker-related downstream tasks.

3. APPROACHES

The overview of our proposed methods is shown in Fig. 1. The details of our proposed methodology are elaborated upon in this section.

3.1. Status Augmentation

Inspired by speed perturbation, a technique commonly used in the field of automatic speech recognition (ASR) [16] and speaker verification (SV) [17] to improve the robustness and generalization of systems, we propose a data augmentation method named "status augmentation". It can account for variations in operational status and increase the robustness of the ASD models, based on the fact that changing the speed of a sound signal can alter its frequency and duration, which are related to the operation status of machines.

During the status augmentation process, the original speech signal is modified by stretching or compressing its duration while maintaining the original pitch or not. This manipulation is achieved by resampling the signal at a different rate or adjusting the playback speed. By applying random

status augmentation to the sound signals, we generate synthetic data that emulates diverse machine operational statuses. These synthetic datasets serve to augment the original training data, enabling the training of more diverse and effective ASD models. Consequently, this augments the models' capacity for generalization.

3.2. Classification Training with Pre-trained Models

The general idea of this paper is to adopt large-scale pre-trained models for fine-tuning on normal machine sounds through classification training, and then detect the anomalous sound based on the fine-tuned models. As illustrated in Figure 1, this process can be segmented into two key stages: training and inference.

During the training process, we fine-tune the pre-trained models on training data with a classifier to categorize the operational status of machines [5]. The input feature for the pre-trained model is a waveform. After traversing through several convolutional blocks and transformer layers, the waveform is encoded into several sequence representations denoted as H_l, where l belongs to the set $\{0, \ldots, L\}$, representing each layer. To enable the model to effectively harness information from different layers, similar to [13], we employ learnable weights w_l to perform a weighted summation of the hidden states, with

$$\widetilde{H} = \sum_{l} w_l \cdot H_l \tag{1}$$

Then, the sequence $\widetilde{H}$ will be aggregated by a pooling layer to generate chunk-level audio embeddings. In our system, the network is optimized to predict the attributes ID from metadata using AAM-softmax [18], as expressed in Equation 2.

$$L_{AAM} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(cos(\theta_{y_i,i}+m))}}{Z} \tag{2}$$

where $Z = e^{s(cos(\theta_{y_i,i}+m))} + \sum_{j=1,j\neq i}^{c} e^{s(cos(\theta_{j,i}))}$, $\theta_{j,i}$ is the angle between the column vector $\mathbf{W_j}$ and embedding $\mathbf{x_i}$, where both $\mathbf{W_j}$ and $\mathbf{x_i}$ are normalized. s is a scaling factor and m is a hyperparameter to control the margin.

During the inference phase, we use the previously fine-tuned network to extract the embeddings from the training set, allowing us to establish the distribution of normal machine sounds. Then, k-NN [19] is adopted as the outlier detector here to compute the anomaly score. This score is then used to ascertain whether a given sound falls within the normal distribution.

3.3. Transformer Pooling

In our observations, we noted that within a recording, the pertinent signal for anomaly detection does not persist continuously. Therefore, it is important for the model to autonomously acquire the skill of discovering effective information. So we chunk the recording into several shorter segments using a sliding window and extract embedding representations from each of these segments. To enhance the aggregation of embeddings from the same recording and capture essential information effectively, we utilize a transformer pooling layer to fuse multiple embeddings into one with the attention mechanism. Consistent with the training objective outlined in Figure 1, we employ a classification to optimize the transformer pooling layer.

4. EXPERIMENTAL SETUP

Dataset We conduct experiments on the development (without additional part) dataset of Task2 in DCASE 2023 Challenge [1], which includes seven machine types (Bearing, Fan, GearBox, Slider, ToyCar, ToyTrain, and Valve). For each machine type, this dataset provides (i) 1000 clips of normal sounds for training, and (ii) 100 clips each of normal and anomalous sounds for the test. Additionally, the attributes of each sample in the training and test data are provided, and we use the attributes as labels for self-supervised classification training [5].

Evaluation Metrics For evaluation metrics, we evaluated the systems with the area under the receiver operating characteristics (ROC) curve (AUC) and the partial AUC (pAUC) following the setup in [1]. We report the harmonic mean (hmean) of the AUC and pAUC scores over all machine types.

Training Configuration For the detailed training configuration, we adopt the AdamW as the optimizer to optimize the whole network. In order to prevent overfitting on training data, we use a relatively small learning rate of $5e$-4 and the weight decay is set to $1e$-4. The whole training process will last 10k steps. We don't apply model selection, but choose the last epoch for evaluation. Besides, to construct the training batch effectively, we randomly sampled 2s from each recording in the training process. All large-scale pre-trained models we used are coming from the huggingface[1]. For back-end k-NN, the distance metric is chosen as cosine distance, and the number of neighbors k is selected as 2, which is same as [20].

5. RESULTS AND ANSLYSIS

The large scale pre-trained models including wav2vec 2.0, UniSpeech, HuBERT and WavLM are pre-trained on distinct datasets with varying training strategies. In our investigation, we assess their impact on the anomalous sound detection task and present the findings in Table 1. For the baseline systems, we choose the reconstruction-based AutoEncoder and classification-based convolutional network MobileNet for comparative analysis. As the results indicate, our fine-tuned systems based on pre-trained models are capable of achieving results comparable to the traditional baseline systems, all

[1] https://huggingface.co/models

Table 1. Results of different large scale pre-trained models on anomalous sound detection of DCASE2023 [1]. The **WS** denotes the weighted sum of latent representation from pre-trained models. **SA** and **TFP** means status augmentation and transformer pooling, respectively. It's noted that all the results we report are the harmonic mean (hmean) of the AUC and pAUC.

WS	SA	TFP	Models	All Hmean	Machines (Hmean)						
					Bearing	Fan	GearBox	Slider	ToyCar	ToyTrain	Valve
-	-	-	AutoEncoder [1]	54.96	61.47	46.55	59.11	57.26	56.55	53.79	52.74
			MobileNet [5]	54.13	60.88	41.29	58.39	96.54	48.32	46.89	52.15
✗	✗	✗	Wav2Vec 2.0	57.89	54.71	56.19	55.08	78.80	50.16	52.99	65.80
			UniSpeech	53.39	53.63	51.52	52.41	58.40	50.49	54.28	53.73
			HuBERT	53.19	53.14	52.79	57.55	55.32	50.64	52.03	51.50
			WavLM	55.30	54.64	53.81	58.34	58.51	53.71	52.85	55.79
			Avg.	54.94	54.03	53.58	55.84	62.76	51.25	53.04	56.70
✓	✗	✗	Wav2Vec 2.0	61.96	61.21	62.24	64.45	74.07	64.40	56.57	54.48
			UniSpeech	61.57	66.89	64.24	62.69	81.04	59.06	54.30	51.18
			HuBERT	62.36	61.83	59.45	**72.90**	81.04	59.75	56.70	53.14
			WavLM	62.33	63.00	52.50	66.98	**83.36**	59.01	58.51	60.92
			Avg.	62.06	63.23	59.61	66.76	79.88	60.55	56.52	54.93
✓	✓	✗	Wav2Vec 2.0	63.17	58.67	60.20	68.46	75.54	64.12	57.38	61.38
			UniSpeech	63.35	63.85	67.95	67.74	72.45	66.64	54.35	55.09
			HuBERT	63.06	61.11	60.91	66.63	72.39	66.66	53.82	63.24
			WavLM	63.13	61.94	60.35	66.73	75.31	58.99	56.78	65.22
			Avg.	63.18	61.39	62.35	67.39	73.92	64.10	55.58	61.23
✓	✓	✓	Wav2Vec 2.0	**64.31**	57.10	62.76	67.52	79.11	63.47	57.35	**67.79**
			UniSpeech	64.30	**65.19**	**69.31**	65.81	75.31	66.41	55.16	57.36
			HuBERT	62.92	59.88	62.65	64.85	72.03	**67.46**	56.04	60.15
			WavLM	63.50	60.21	54.80	69.06	71.05	66.49	**61.89**	63.99
			Avg.	63.76	60.60	62.38	66.81	74.38	65.96	57.61	62.32

without using WS, SA, TFP strategies. This underscores the efficacy of the acoustic features extracted by pre-trained models in the context of ASD tasks, even though their initial pre-training objective pertains to speech-related tasks.

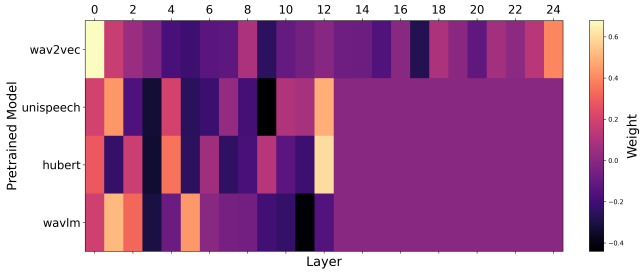

Fig. 2. The visualization of weight value w_i for each layer. It should be noted that the wav2vec 2.0 has 24 layers and others only have 12 layers.

Instead of simply using the outputs from the last layer of the pre-trained models, we employ a weighted sum (WS) approach to blend the hidden states from various layers, thereby merging acoustic features across different scales. According to the results in Table 1, pre-trained models with weighted sum obtain significant performance improvement (Avg. All Hmean from 54.94 to 62.06) and outperform the baselines by a large margin. In addition, we provide a visualization of the weight vale w_i for each layer in Fig. 2. The figure shows that the distribution of layers with high importance is relatively uniform, and there is no phenomenon of concentration in shallow or deep layers like [13]. It indicates that the anomalous information may be hidden at different scales, and this multi-layer fusion mechanism can effectively improve the robustness of ASD task.

Based on the pre-trained models, we also evaluate the effectiveness of the status augmentation (SA) we proposed in Table 1. It can be observed that there exists an improvement in consistency across all pre-trained models. Similarly, we also provide the comparison results of transformer pooling (TFP). Performance will also be further improved with TFP.

6. CONCLUSION

In this paper, to tackle the poor generalization problem in ASD caused by the challenges of data collection and complex acoustic environment, we explore several pre-trained models which are trained on large scale speech data for robust performance. Comparing with traditional baseline systems, it can achieve excellent results on DCASE2023 dataset. In addition, we propose status augmentation for augmenting normal sounds. Finally, we adopt a transformer based pooling method to gather the effective information from the recordings. With these strategies, we outperform the commonly used methods by a large margin and achieve 2nd place in Task2 of DCASE2023.

7. REFERENCES

[1] Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, and Yohei Kawaguchi, "Description and discussion on dcase 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2305.07828*, 2023.

[2] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, and Masahiro Yasuda, "First-shot anomaly detection for machine condition monitoring: a domain generalization baseline," *arXiv preprint arXiv:2303.00455*, 2023.

[3] Kota Dohi, Takashi Endo, Harsh Purohit, Ryo Tanabe, and Yohei Kawaguchi, "Flow-based self-supervised density estimation for anomalous sound detection," in *Proc. ICASSP*. IEEE, 2021, pp. 336–340.

[4] Kaori Suefusa, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, and Yohei Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. ICASSP*. IEEE, 2020, pp. 271–275.

[5] Ritwik Giri, Srikanth V Tenneti, Fangzhou Cheng, Karim Helwani, Umut Isik, and Arvindh Krishnaswamy, "Self-supervised classification for detecting anomalous sounds," *DCASE*, 2020.

[6] Anbai Jiang, Wei-Qiang Zhang, Yufeng Deng, Pingyi Fan, and Jia Liu, "Unsupervised anomaly detection and localization of machine audio: A gan-based approach," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[7] Han Chen, Yan Song, Li-Rong Dai, Ian McLoughlin, and Lin Liu, "Self-supervised representation learning for unsupervised anomalous sound detection under domain shift," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 471–475.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NIPS*, vol. 33, pp. 12449–12460, 2020.

[10] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. ASLP.*, vol. 29, pp. 3451–3460, 2021.

[11] Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang, "Unispeech: Unified speech representation learning with labeled and unlabeled data," in *Proc. ICML*. PMLR, 2021, pp. 10937–10947.

[12] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE JSTSP*, vol. 16, no. 6, pp. 1505–1518, 2022.

[13] Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *Proc. ICASSP*. IEEE, 2022, pp. 6147–6151.

[14] Liang Xu, Lizhong Wang, Sijun Bi, Hanyue Liu, and Jing Wang, "Semi-supervised sound event detection with pre-trained model," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[15] Alexei Baevski, Steffen Schneider, and Michael Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[16] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[17] Zhengyang Chen, Bing Han, Xu Xiang, Houjun Huang, Bei Liu, and Yanmin Qian, "Build a sre challenge system: Lessons from voxsrc 2022 and cnsrc 2022," *arXiv preprint arXiv:2211.00815*, 2022.

[18] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4690–4699.

[19] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim, "Efficient algorithms for mining outliers from large data sets," in *Proc. SIGMOD*, 2000, pp. 427–438.

[20] Anbai Jiang, Qijun Hou, Jia Liu, Pingyi Fan, Jitao Ma, Cheng Lu, Yuanzhi Zhai, Yufeng Deng, and Wei-Qiang Zhang, "Thuee system for first-shot unsupervised anomalous sound detection for machine condition monitoring," *DCASE*, 2023.