INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Anomalous Sound Detection Using Self-Attention-Based Frequency Pattern Analysis of Machine Sounds

Hejing Zhang[1], *Jian Guan*[1,*], *Qiaoxi Zhu*[2], *Feiyang Xiao*[1], *Youde Liu*[3]

[1] Group of Intelligent Signal Processing, Harbin Engineering University, China
[2] Centre for Audio, Acoustics and Vibration, University of Technology Sydney, Australia
[3] School of Computer Science and Technology, Harbin Institute of Technology, China

{zhanghejing, j.guan}@hrbeu.edu.cn; qiaoxi.zhu@gmail.com; xiaofeiyang128@gmail.com; liuyoudedl@163.com

Abstract

Different machines can exhibit diverse frequency patterns in their emitted sound. This feature has been recently explored in anomaly sound detection and reached state-of-the-art performance. However, existing methods rely on the manual or empirical determination of the frequency filter by observing the effective frequency range in the training data, which may be impractical for general application. This paper proposes an anomalous sound detection method using self-attention-based frequency pattern analysis and spectral-temporal information fusion. Our experiments demonstrate that the self-attention module automatically and adaptively analyses the effective frequencies of a machine sound and enhances that information in the spectral feature representation. With spectral-temporal information fusion, the obtained audio feature eventually improves the anomaly detection performance on the DCASE 2020 Challenge Task 2 dataset.

Index Terms: Anomalous sound detection, frequency pattern analysis, self-attention, feature representation

1. Introduction

Anomalous sound detection (ASD) aims to automatically determine whether the state of a target object is normal or anomalous by analyzing the sound emitted by the object [1–7]. ASD is commonly an unsupervised task due to the infrequent and varied occurrence of anomalous machine sounds in real-world scenarios [1, 3, 5–8]. Therefore, only normal sounds are employed for training to learn the audio feature distribution of normal sounds. The distance between the test sound and the learned normal sound distribution is calculated to detect the anomalous sound having a distance value larger than a threshold [9–11].

As an unsupervised task, ASD learns the feature of normal sounds to detect anomalous sounds. If the learnt feature also fits with the anomalous sound, the effectiveness of anomaly detection could be limited. Log-Mel spectrogram has been widely used as the input feature of the machine sound in ASD methods, such as [4,12–14]. However, in our previous work [8], we found that using Log-Mel spectrogram as the audio feature can be ineffective in distinguishing normal and anomalies, as it might filter out high-frequency components of anomaly sound, where distinct features may exist. So spectral-temporal information fusion (STgram) as the audio feature was proposed in [8], utilising both the Log-Mel spectrogram and temporal feature extracted from machine sounds. Using STgram, the STgram-MFN method was developed for ASD [8], which achieved state-of-the-art performance on the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge 2020 Task 2 dataset.

Further investigation indicates that some machine types exhibit prominent characteristics in high frequencies, as evidenced by analyzing the spectrum of machine sounds [3, 5]. Additionally, the models used for anomaly detection in ASD rely heavily on higher frequencies to distinguish between normal and abnormal sounds [3]. To obtain the audio feature, a high-pass filter is applied before passing it through the Mel filter in [5], which ranked top 1 in DCASE 2022 Challenge Task 2. The results of experiments demonstrate that this pre-processed feature improved anomaly detection for several machine types, including ToyCar, ToyTrain, Gearbox, and Valve. However, this pre-processing technique [3, 5] relies on the manual or empirical determination of the high-pass filter by observing the effective frequency range in the training data. This approach may be imprecise or time-consuming when implementing ASD in real-world settings.

In general, various machines can exhibit diverse frequency patterns in their emitted sound, and normal or abnormal sounds can also possess distinct frequency patterns. In practical applications, it is strongly desired for ASD to have the capability to automatically identify the frequency pattern of a machine sound and adjust its processing accordingly based on the specific frequency pattern to attain the most optimal results.

In this paper, we propose an ASD method using self-attention-based frequency pattern analysis (ASD-AFPA) to extract essential information over frequencies of the machine sound for improved anomaly detection. It uses STgram-MFN [8] as the backbone. However, ASD-AFPA differs from STgram-MFN in that it integrates self-attention mechanism [15] after the Log-Mel converter to achieve a more effective spectral feature, before performing spectral-temporal information fusion. To the best of our knowledge, the proposed method is the first to introduce automatic frequency pattern analysis for anomalous sound detection. Our experiments demonstrate that the self-attention module automatically and adaptively analyses the effective frequencies of a machine sound for ASD and enhances that information in the spectral feature representation, which eventually improves the audio feature obtained from the spectral-temporal information fusion.

2. Proposed Method

The proposed ASD method using self-attention-based frequency pattern analysis (ASD-AFPA) is illustrated in Figure 1. The method is built on our previous spectral-temporal information fusion-based ASD approach, STgram-MFN [8], but with spectral features boosted using a multi-head self-attention mechanism [15]. This enables the method to automatically de-

*Corresponding author

10.21437/Interspeech.2023-2416

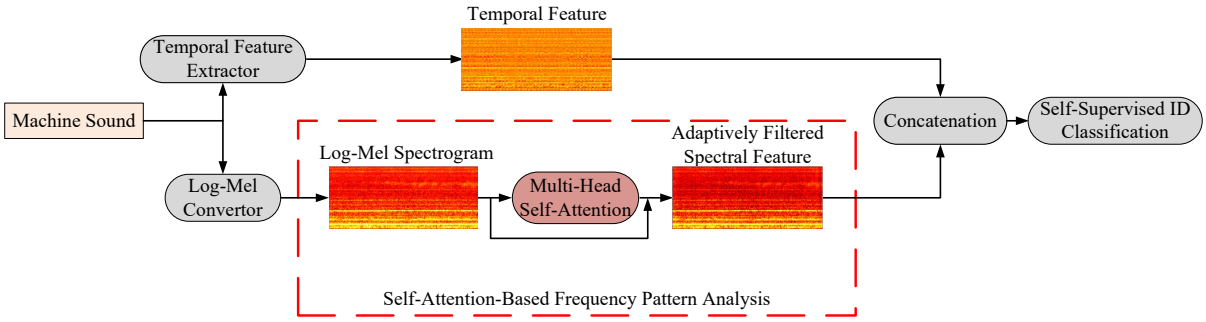


Figure 1: *Framework of the proposed method with self-attention-based frequency pattern analysis.*

tect the frequency pattern of the machine sound and modify its processing accordingly, thereby achieving the most optimal performance for ASD, as demonstrated by the experiment in Section 3. Here, Section 2.1 introduces the STgram-MFN method as the backbone of the proposed method, and Section 2.2 details the multi-head self-attention mechanism enabling the frequency pattern analysis of the machine sound.

2.1. Spectral-Temporal Feature Fusion

For a machine sound as single-channel audio signal $\mathbf{x} \in \mathbb{R}^{1\times L}$ with length L, its Log-Mel spectrogram is $\mathbf{X}_F \in \mathbb{R}^{M\times N}$, where M denotes the Mel bins of the Log-Mel spectrogram (i.e., the number of frequency components) and N denotes the number of time frames. The temporal feature is obtained as

$$\mathbf{X}_T = TN(\mathbf{x}), \tag{1}$$

where $TN(\cdot)$ represents TgramNet in [8], and $\mathbf{X}_T \in \mathbb{R}^{M\times N}$ has the same dimension with $\mathbf{X}_F$. The audio representation $\mathbf{X} \in \mathbb{R}^{2\times M\times N}$ through spectral-temporal feature fusion is

$$\mathbf{X} = Concat_{3D}(\mathbf{X}_F, \mathbf{X}_T), \tag{2}$$

where $Concat_{3D}(\cdot)$ is a 3-dimensional concatenation operation.

After the feature concatenation, the audio representation $\mathbf{X}$ will be passed to a self-supervised ID classification for anomaly detection, including MobileFaceNet [16] (MFN) as the classifier and ArcFace [17] as the loss function, which is conducive to enhancing inter-class compactness and amplifying intra-class differences.

Our prior research [8] has shown that combining spectral and temporal features enhances audio representation for ASD. The proposed method in this paper retains this structure to preserve this benefit.

2.2. Self-Attention-Based Frequency Pattern Analysis

The proposed method ASD-AFPA applies the multi-head self-attention (MHSA) mechanism [15] to the Log-Mel spectrogram $\mathbf{X}_F$ to automatically analyse the frequency pattern of the machine sound and adaptively using the obtained information of effective frequency components for optimised ASD.

First, to prevent information interference between time frames and allows better analysis of the effective frequency components from the Log-Mel spectrogram, we segment the Log-Mel spectrogram over the time dimension as

$$\mathbf{X}_F = [\mathbf{X}_F(1), \cdots, \mathbf{X}_F(i), \cdots, \mathbf{X}_F(I)], \tag{3}$$

where $\mathbf{X}_F(i) \in \mathbb{R}^{M\times n}, i = 1, 2, \cdots, I$, and $n = N/I$. Here, I is the number of heads of the multi-head self-attention used for frequency pattern analysis.

The latent parameters $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{M\times N}$ of the multi-head self-attention mechanism are obtained by the linear mapping of the input Log-Mel spectrogram $\mathbf{X}_F$, and they are used to calculate the weights of different frequency components. $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are calculated as

$$\begin{cases} \mathbf{Q} = \mathbf{X}_F \cdot \mathbf{W}_Q, \\ \mathbf{K} = \mathbf{X}_F \cdot \mathbf{W}_K, \\ \mathbf{V} = \mathbf{X}_F \cdot \mathbf{W}_V, \end{cases} \tag{4}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{N\times N}$ are learnable parameter matrices, and their function is to linearly map the input $\mathbf{X}_F$.

To achieve the proposed frequency pattern analysis and enhance the effective frequency components information in the audio feature, ASD-AFPA uses a self-attention mechanism for each part of the segmented input Log-Mel spectrogram to obtain the frequency patterns, which can be calculated as

$$A(\mathbf{X}_F(i)) = softmax\left(\frac{\mathbf{Q}_i \cdot \mathbf{K}_i^\top}{\sqrt{n}}\right) \cdot \mathbf{V}_i, \tag{5}$$

$$\mathbf{D}i = softmax\left(\frac{\mathbf{Q}_i \cdot \mathbf{K}_i^\top}{\sqrt{n}}\right), \tag{6}$$

where $\top$ represents the transposition of the matrix, and n represents the dimension of the time frame of $\mathbf{X}_F(i)$. $\mathbf{Q}_i$, $\mathbf{K}_i$, and $\mathbf{V}_i$ are part of the latent parameters $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, respectively, corresponding to the spectrogram segment $\mathbf{X}_F(i)$. Here, $A(\mathbf{X}_F(i))$ denotes the output of the self-attention applied on $\mathbf{X}_F(i)$, and $\mathbf{D}i \in \mathbb{R}^{M\times M}$ is the frequency pattern weight matrix (i.e., attention map) of $\mathbf{X}_F(i)$ learned from the self-attention mechanism. Note that the values in $\mathbf{D}i$ range from 0 to 1 and represent the importance of the frequency components in the Log-Mel spectrogram. A larger value indicates the weighted frequency components containing more effective information.

The output of MHSA is $MHSA(\mathbf{X}_F)$, which is obtained by passing $\mathbf{X}_F(i)(i = 1, ..., I)$ through the self-attention mechanism and connecting $A(\mathbf{X}_F(i))$ on the time frame dimension, that

$$MHSA(\mathbf{X}_F) = concat(A(\mathbf{X}_F(1)), \cdots, A(\mathbf{X}_F(I))). \tag{7}$$

To obtain the important information in frequency components while preserving the global information of the Log-Mel spectrogram, we add the residual to the output of MHSA

Table 1: *Performance on AUC (%) and pAUC (%) for different machine types. STgram-MFN is the backbone of the proposed ASD-AFPA method. The proposed method only differs from the backbone in adding the multi-head self-attention to adaptively learn the important frequency patterns for more effective audio feature learning for ASD.*

Methods	Fan		Pump		Slider		Valve		ToyCar		ToyConveyor		Average	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
IDNN [4]	67.71	52.90	73.76	61.07	86.45	67.58	84.09	64.94	78.69	69.22	71.07	59.70	76.96	62.57
MobileNetV2 [13]	80.19	74.40	82.53	76.50	95.27	85.22	88.65	87.98	87.66	85.92	69.71	56.43	84.34	77.74
Glow_Aff [14]	74.90	65.30	83.40	73.80	94.60	82.80	91.40	75.00	92.20	84.10	71.50	59.00	85.20	73.90
STgram-MFN [8] (backbone)	94.04	88.97	91.94	81.75	99.55	97.61	**99.64**	**98.44**	94.44	87.68	74.57	63.60	92.36	86.34
ASD-AFPA	**97.55**	**93.48**	**94.46**	**86.76**	**99.69**	**98.40**	99.12	95.42	**96.15**	**89.45**	**76.49**	**64.21**	**93.91**	**87.95**

$MHSA(\mathbf{X}_F)$. The audio feature $\hat{\mathbf{X}}_F \in \mathbb{R}^{M\times N}$ with adaptively frequency pattern analysis can be calculated as

$$\hat{\mathbf{X}}_F = MHSA(\mathbf{X}_F) + \mathbf{X}_F. \tag{8}$$

Finally, the enhanced audio feature representation $\hat{\mathbf{X}} \in \mathbb{R}^{2\times M\times N}$ can be obtained by fusing the adaptively filtered spectral feature $\hat{\mathbf{X}}_F$ and the temporal feature $\mathbf{X}_T$, that

$$\hat{\mathbf{X}} = Concat_{3D}(\hat{\mathbf{X}}_F, \mathbf{X}_T). \tag{9}$$

We adopt the temporal feature to compensate for the possibly missed information in the Log-Mel spectrogram, further improving the audio feature representation with the adaptive frequency pattern analysis.

3. Experimental Results

To assess the effectiveness of the proposed method, we performed experiments on the DCASE 2020 Challenge Task 2 dataset [1]. The experimental results demonstrated that incorporating self-attention-based frequency pattern analysis into the existing backbone improved ASD performance compared to state-of-the-art techniques. Furthermore, the ablation study validated the improvement from the proposed self-attention-based frequency pattern analysis. We also present illustrative examples to showcase the important frequency components detected from machine sounds and the resulting changes in the learned spectral features, ultimately leading to improved performance in detecting anomalous sounds.

3.1. Experimental Setup

Dataset: We evaluated our proposed method on the DCASE 2020 Challenge Task 2 dataset [1]. The dataset consists of six machine types (Fan, Pump, Slider, Valve, ToyCar, and ToyConveyor), each comprising sounds from four different machine IDs, except for ToyConveyor, which has three different machine IDs. The development and additional datasets' training data is used for training, and the development dataset's test data is used for evaluation. We didn't choose datasets from DCASE 2021 [18] or 2022 [19] since they focus on the domain shift problem, which is out of the scope of this paper.

Evaluation metrics: The evaluation metrics include the area under the receiver operating characteristic curve (AUC) and the partial-AUC (pAUC), following [4, 8, 13, 14], where pAUC represents the AUC over a low false-positive-rate range $[0, 0.1]$ [1]. A larger metric value indicates a better distinguishing ability for anomalous sound detection.

Parameter settings: We employ Adam optimizer [20] with a learning rate of 1×10^{-4} for model training by 200 epochs, and cosine annealing is applied for learning rate decay. The margin and scale of the ArcFace loss [17] are empirically set as 1.0 and 30, respectively. The number of heads in Eq. (3) is 6. The number of the Mel bins (i.e., frequency components) of the input Log-Mel spectrogram is 128, with 312 time frames.

3.2. Performance Comparison and Ablation Study

The proposed method ASD-AFPA is compared with the state-of-the-art methods on the DCASE 2020 dataset, IDNN [4], MobileNetV2 [13], Glow_Aff [14], and STgram-MFN [8]. Table 1 shows that the proposed ASD-AFPA method significantly improves the ASD performance for all machine types (except Valve), with 1.55% improvement on AUC and 1.61% improvement on pAUC, averaged over all the six machine types, compared with STgram-MFN [8] that achieved the best performance amongst other methods. Note that the STgram-MFN is the backbone of the proposed method, and the only difference between these two methods is that the backbone does not have the self-attention-based frequency pattern analysis, but the proposed ASD-AFPA method does. So this result demonstrated that the proposed method's adaptive frequency pattern analysis (AFPA) is effective for ASD.

3.3. Visualisation Analysis

Figure 2 presents illustrative examples to show the important frequency components detected from machine sounds and the resulting changes in the learned spectral features, ultimately leading to improved performance in detecting anomalous sounds. Figure 2 includes examples of machine sound from different machine individuals of the same machine type (i.e., ID 00 and ID 04 for Fan), as well as the normal and anomaly sound of the same machine. Column (a) presents the Log-Mel spectrograms of the sound signals. Column (b) presents the learnt frequency patterns, i.e., the mean pooling results of all the learnt frequency weight matrices from Eq. (6). Column (c) presents the learnt audio feature with self-attention-based frequency pattern analysis.

From Figure 2, we can see that our method can learn different frequency patterns for different operating sound signals, as demonstrated in the blue boxes in column (b), which can highlight the important frequency components of the input Log-Mel spectrograms, by giving large weights to these frequency components, where the important frequency components with effective information can be highlighted, as illustrated in column (c). In addition, we can see that our proposed method is effective for all test sound signals, regardless of machine conditions, including signals from different machines (i.e., ID 00 and ID 04 of machine type Fan) and the machine operating status (i.e., normal or abnormal). The results further verify the effectiveness of the proposed method, and show how our method can achieve the adaptive frequency pattern analysis, thus obtaining enhanced audio feature representation to improve the detection

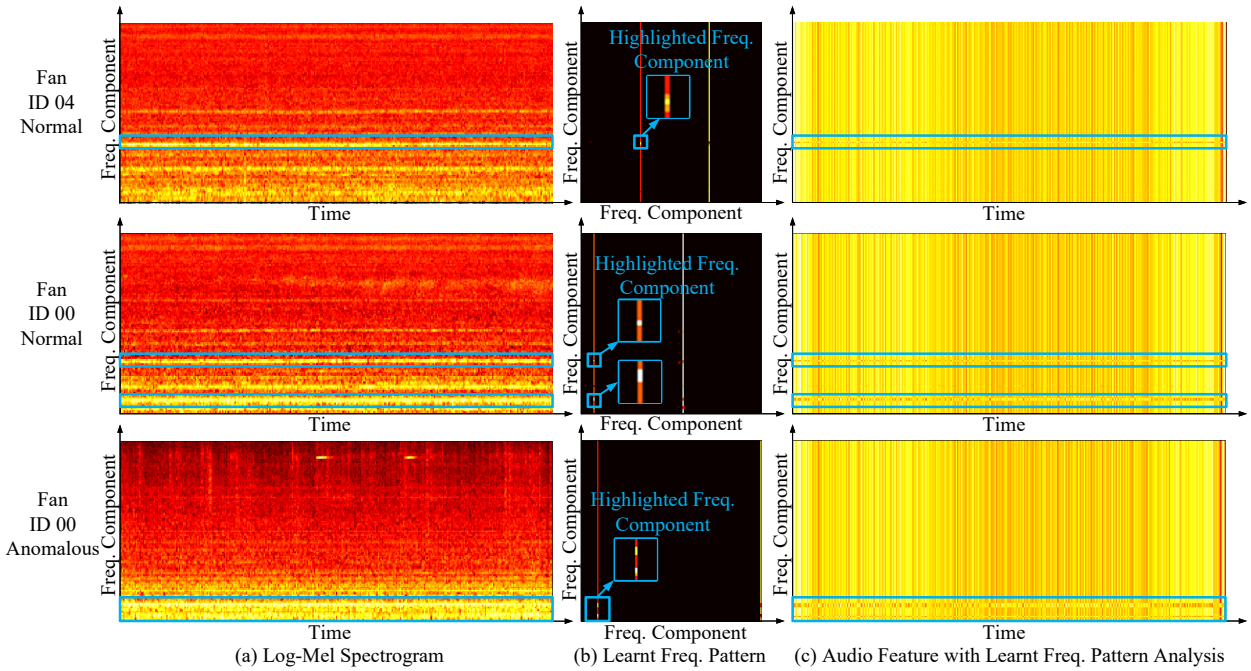
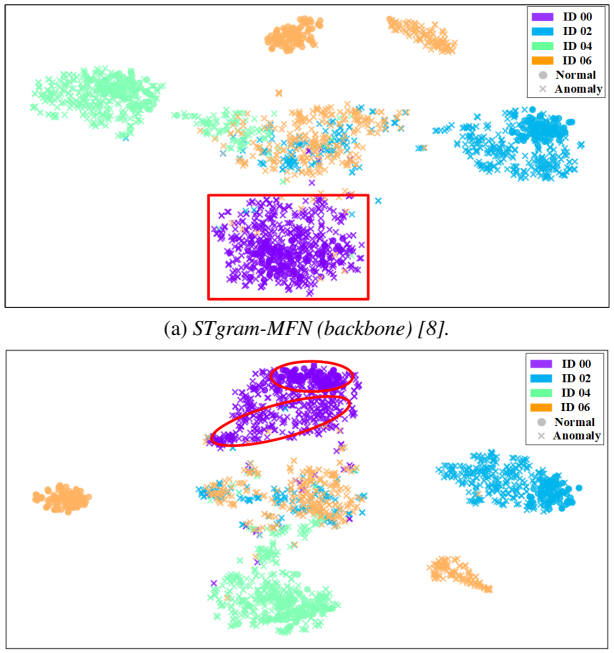


Figure 2: *Illustration of the audio feature with the learnt frequency pattern for machine type Fan with ID 04 and 00, respectively. (a) The Log-Mel spectrograms of the input audio signals; (b) The learnt frequency patterns, i.e., the learnt weight matrices corresponding to the input Log-Mel spectrograms of (a); (c) The learnt audio feature with the proposed self-attention-based frequency pattern analysis. For better understanding, we mark the highlighted important frequency components obtained by the learnt frequency patterns, as shown in the blue boxes of columns (b) and (c).*



(a) *STgram-MFN (backbone) [8].*

(b) *The proposed ASD-AFPA method.*

Figure 3: *The t-SNE visualisation comparison between the backbone method STgram-MFN [8] and the proposed ASD-AFPA for the machine type Fan.*

performance.

In addition, we provide the t-SNE visualisation comparison between the backbone STgram-MFN [8] and our proposed ASD-AFPA method for the machine type Fan. As illustrated in Figure 3, where we can see that our method with the adaptive frequency pattern analysis can further improve the distinguishing ability for anomalous sound detection. The normal and anomalous sound of machine ID 00 can be better distinguished than the backbone method, which further verifies the effectiveness of the proposed method.

4. Conclusion

In this paper, we propose an anomalous sound detection method using self-attention-based frequency pattern analysis. It enables automatic detection of the individual frequency pattern of a machine sound and enhances the spectral-temporal feature fusion-based audio feature representation for anomaly detection. Experiments have shown that our proposed approach outperforms existing methods and can identify important frequency components that contribute to enhanced performance in detecting anomalous sounds.

5. Acknowledgements

This work was partly supported by the Natural Science Foundation of Heilongjiang Province under Grant No. YQ2020F010, and a GHfund with Grant No. 202302026860.

6. References

[1] 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 task 2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. of the 5th Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 81–85.

[2] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE2021 challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in *Proc. of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 186–190.

[3] K. T. Mai, T. Davies, L. D. Griffin, and E. Benetos, "Explaining the decision of anomalous sound detectors," in *Proc. of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

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

[5] Y. Zeng, H. Liu, L. Xu, Y. Zhou, and L. Gan, "Robust anomaly sound detection framework for machine condition monitoring," DCASE2022 Challenge, Tech. Rep, Tech. Rep., 2022.

[6] Y. Park and I. D. Yun, "Fast adaptive RNN encoder–decoder for anomaly detection in SMD assembly machine," *Sensors*, vol. 18, no. 10, p. 3573, 2018.

[7] D. Y. Oh and I. D. Yun, "Residual error based anomaly detection using auto-encoder in SMD machine sound," *Sensors*, vol. 18, no. 5, p. 1308, 2018.

[8] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 816–820.

[9] F. Xiao, Y. Liu, Y. Wei, J. Guan, Q. Zhu, T. Zheng, and J. Han, "The DCASE2022 challenge task 2 system: Anomalous sound detection with self-supervised attribute classification and gmm-based clustering," DCASE2022 Challenge, Tech. Rep., July 2022.

[10] I. Kuroyanagi, T. Hayashi, K. Takeda, and T. Toda, "Two-stage anomalous sound detection systems using domain generalization and specialization techniques," DCASE2022 Challenge, Tech. Rep., July 2022.

[11] Y. Wei, J. Guan, H. Lan, and W. Wang, "Anomalous sound detection system with self-challenge and metric evaluation for DCASE2022 challenge task 2," DCASE2022 Challenge, Tech. Rep., July 2022.

[12] S. Kapka, "ID-conditioned auto-encoder for unsupervised anomaly detection," in *Proc. of the 5th Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 71–75.

[13] R. Giri, S. V. Tenneti, F. Cheng, K. Helwani, U. Isik, and A. Krishnaswamy, "Self-supervised classification for detecting anomalous sounds," in *Proc. of the 5th Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 46–50.

[14] K. Dohi, T. Endo, H. Purohit, R. Tanabe, and Y. Kawaguchi, "Flow-based self-supervised density estimation for anomalous sound detection," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 336–340.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, vol. 30. Curran Associates, Inc., 2017.

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

[17] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[18] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," *In arXiv e-prints: 2106.04492, 1–5*, 2021.

[19] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on DCASE2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," *arXiv e-prints: 2206.05876*, 2022.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of International Conference on Learning Representations (ICLR)*, 2014.