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

Hejing Zhang1, Jian Guan1*, Qiaoxi Zhu2, Feiyang Xiao1, Youde Liu3

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; xiaofeyang128@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-
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 $x \in \mathbb{R}^{1 \times L}$ with length $L$, its Log-Mel spectrogram is $X_F \in \mathbb{R}^{M \times N}$, 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

$$X_T = T N(x),$$

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

$$X = Concat_3D(X_F, X_T),$$

where $Concat_3D(\cdot)$ is a 3-dimensional concatenation operation.

After the feature concatenation, the audio representation $X$ 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 $X_F$ 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

$$X_F = [X_F(1), \cdots, X_F(i), \cdots, X_F(I)],$$

where $X_F(i) \in \mathbb{R}^{M \times n}, i = 1, 2, \cdots, I$, 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 $Q, K, V \in \mathbb{R}^{M \times N}$ of the multi-head self-attention mechanism are obtained by the linear mapping of the input Log-Mel spectrogram $X_F$, and they are used to calculate the weights of different frequency components. $Q, K, V$ are calculated as

$$Q = X_F \cdot W_Q, \quad K = X_F \cdot W_K, \quad V = X_F \cdot W_V,$$

where $W_Q, W_K, W_V \in \mathbb{R}^{N \times N}$ are learnable parameter matrices, and their function is to linearly map the input $X_F$.

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

$$A(X_F(i)) = \text{softmax} \left( \frac{Q \cdot K^T}{\sqrt{n}} \right) \cdot V,$$

where $\cdot$ represents the transposition of the matrix, and $n$ represents the dimension of the time frame of $X_F(i)$. $Q$, $K$, and $V$ are part of the latent parameters $Q$, $K$, and $V$, respectively, corresponding to the spectrogram segment $X_F(i)$. Here, $A(X_F(i))$ denotes the output of the self-attention applied on $X_F(i)$, and $D_i \in \mathbb{R}^{M \times M}$ is the frequency pattern weight matrix (i.e., attention map) of $X_F(i)$ learned from the self-attention mechanism. Note that the values in $D_i$ 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 $MHSA(X_F)$, which is obtained by passing $X_F(i)(i = 1, \cdots, I)$ through the self-attention mechanism and connecting $A(X_F(i))$ on the time frame dimension, that

$$MHSA(X_F) = \text{concat}(A(X_F(1)), \cdots, A(X_F(I))).$$

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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Fan</th>
<th>Pump</th>
<th>Slider</th>
<th>Valve</th>
<th>ToyCar</th>
<th>ToyConveyor</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>pAUC</td>
<td>AUC</td>
<td>pAUC</td>
<td>AUC</td>
<td>pAUC</td>
<td>AUC</td>
</tr>
<tr>
<td>IDNN [4]</td>
<td>67.7</td>
<td>52.90</td>
<td>71.76</td>
<td>61.92</td>
<td>86.45</td>
<td>67.58</td>
<td>84.09</td>
</tr>
<tr>
<td>MobileNetV2 [13]</td>
<td>80.19</td>
<td>74.40</td>
<td>82.53</td>
<td>76.50</td>
<td>95.27</td>
<td>85.22</td>
<td>88.65</td>
</tr>
<tr>
<td>Glow_Aff [14]</td>
<td>74.90</td>
<td>65.30</td>
<td>83.40</td>
<td>73.80</td>
<td>94.60</td>
<td>82.80</td>
<td>91.40</td>
</tr>
</tbody>
</table>

STgram-MFN [8] (backbone) | 94.04| 88.97| 91.94  | 81.75 | 99.55  | 97.61       | 99.64   |

ASD-AFPA | 97.55| 93.48| 94.46  | 86.76 | 99.69  | 98.40       | 99.12   |

Note: 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 (APPA) is effective for ASD.

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 (APPA) 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
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


