ANOMALOUS SOUND DETECTION USING SPECTRAL-TEMPORAL INFORMATION FUSION

Youde Liu[1], *Jian Guan*[1*], *Qiaoxi Zhu*[2], *Wenwu Wang*[3]

[1]Group of Intelligent Signal Processing, College of Computer Science and Technology
Harbin Engineering University, China
[2]Centre for Audio, Acoustics and Vibration, University of Technology Sydney, Australia
[3]Centre for Vision Speech and Signal Processing, University of Surrey, UK

ABSTRACT

Unsupervised anomalous sound detection aims to detect unknown abnormal sounds of machines from normal sounds. However, the state-of-the-art approaches are not always stable and perform dramatically differently even for machines of the same type, making it impractical for general applications. This paper proposes a spectral-temporal fusion based self-supervised method to model the feature of the normal sound, which improves the stability and performance consistency in detection of anomalous sounds from individual machines, even of the same type. Experiments on the DCASE 2020 Challenge Task 2 dataset show that the proposed method achieved 81.39%, 83.48%, 98.22% and 98.83% in terms of the minimum AUC (worst-case detection performance amongst individuals) in four types of real machines (fan, pump, slider and valve), respectively, giving 31.79%, 17.78%, 10.42% and 21.13% improvement compared to the state-of-the-art method, i.e., Glow_Aff. Moreover, the proposed method has improved AUC (average performance of individuals) for all the types of machines in the dataset. The source codes are available at https://github.com/liuyoude/STgram_MFN

Index Terms— Anomalous sound detection, feature fusion, self-supervised learning

1. INTRODUCTION

Anomalous sound detection (ASD) aims to automatically identify whether a target object (e.g., a machine or a device) is normal or abnormal from the sound emitted. Collecting anomalous sound data is not a trivial task due to their diversity and scarcity in the real world. Therefore, normal sounds are often used to learn the features of the normal sound, and these learnt features are then used to distinguish the normal and abnormal sound [1–3]. Conventional ASD systems have used autoencoder (AE), such as interpolation deep neural network (IDNN) [4] and ID-conditioned autoencoder [5]. They learn the feature of normal sounds by minimizing the reconstruction error and using the reconstruction error as the score to detect the anomalies. However, since the training procedure does not involve the anomalous sound, the effectiveness of such model could be limited if the trained feature also fits with the anomalous sound [6].

To better model the normal sound feature, the self-supervised classification approach has been proposed by using the metadata, i.e., machine type and machine identity (ID) in addition to condition (normal/anomaly), accompanying the audio files [7], and it performs better than the AE-based unsupervised methods. However, this method is not always stable and performs differently even for machines of the same type [8]. As a solution, the flow-based self-supervised density estimation (i.e., Glow_Aff) was proposed in [8], using normalizing flow such as generative flow (Glow) [9] or masked autoregressive flow (MAF) [10]. This method improves the detection performance on one machine ID by introducing an auxiliary task to distinguish the sound data of that machine ID (target data) from sound data of other machine IDs with the same machine type (outlier data). Thus, this approach requires different training models for different machine IDs of each machine type, which is not desired for general applications. In addition, the detection stability of this approach on sounds of individual machines of the same type is still limited.

To develop a general method with stability and avoid specialized training for each machine ID, we study empirically the effectiveness of features for ASD, such as the log-Mel spectrogram which has been widely used as input feature in ASD [4, 5, 7, 8]. This feature was designed based on human auditory perception, using mel filter bank to capture the information in various frequencies [11]. However, it might filter out high-frequency components of anomaly sound, where distinct features may exist. Thus, the log-Mel spectrogram feature may not be able to fully distinguish the normal and anomalous sounds, resulting in unstable performance when

*Corresponding Author

This work was partly supported by the Natural Science Foundation of Heilongjiang Province under Grant No. YQ2020F010, and a Newton Institutional Links Award from the British Council with Grant No. 623805725.

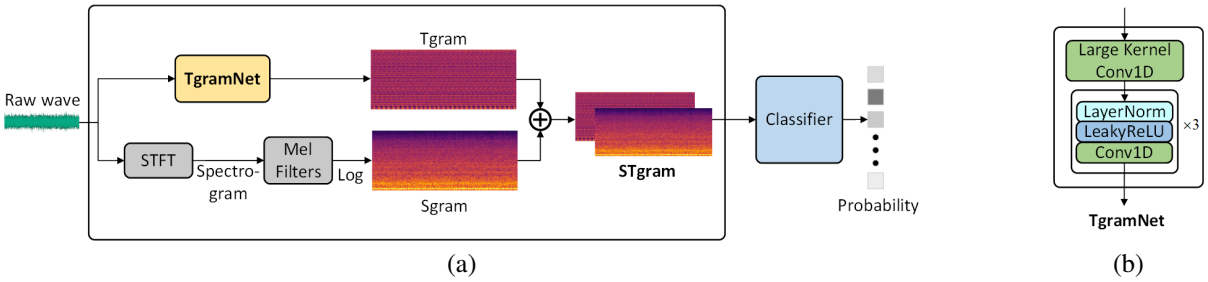


Fig. 1. The framework of the proposed method for anomalous sound detection. (a) The spectral-temporal feature is extracted from the raw wave through a CNN-based network (TgramNet) for the temporal feature and the log-Mel spectrogram for the frequency feature. $\oplus$ denotes concatenation operation for feature fusion. The spectral-temporal feature is then fed into the classifier for anomalous sound detection. (b) Details in TgramNet.

used with the self-supervised approaches. There is potential to use temporal information to complement the log-Mel spectrogram, which is recently studied in audio pattern recognition [12]. However, the incorporation of temporal information has not been reported for ASD.

In this paper, a spectral-temporal feature, STgram, is proposed as the input feature for self-supervised classification approach by fusing the log-Mel spectrogram (Sgram) and the temporal feature (Tgram). Here, the temporal feature is extracted from the raw wave by a proposed CNN-based network (TgramNet) to compensate for the anomalous information unavailable from the log-Mel spectrogram. The spectral-temporal feature is then fed into the classifier, i.e., MobileFaceNet (MFN) [13], to learn the delicate feature representation of normal sounds for the detection of anomalies.

The proposed method is evaluated by experiments on the DCASE 2020 Challenge Task 2 dataset [3], as compared with the state-of-the-art methods, over six types of machine sound. The proposed method improves AUC and pAUC performance and largely increases performance stability observed from the mAUC metric. Furthermore, the ablation study presents the comparison with potential alternatives using temporal only or spectral-temporal input feature, to show that the proposed method provides an effective way of utilizing both the spectral and temporal features for anomalous sound detection.

2. PROPOSED METHOD

This section presents the proposed method for anomalous sound detection in detail. The overall framework of the proposed method is given in Fig. 1.

2.1. Spectral-temporal feature fusion

Let $\boldsymbol{x} \in \mathbb{R}^{1\times L}$ be the input single-channel audio signal with the length L. The log-Mel spectrogram of $\boldsymbol{x}$ is $\boldsymbol{F}_S \in \mathbb{R}^{M\times N}$, where M denotes the dimension of the spectrum feature (i.e., number of Mel bins) and N is the number of time frames. The

Table 1. The architecture of TgramNet.

Operation	c	k	s	p	n
Conv1D	M	W	H	$W/2$	$\times 1$
LayerNorm	-	-	-	-	
LeakyReLU	-	-	-	-	$\times 3$
Conv1D	M	3	1	1	

* Here, c, k, s, p, n represent number of channels, kernel size, stride, padding size and number of layers, respectively. M is the number of Mel bins. W and H are the window size and hop length of STFT, respectively.

log-Mel spectrogram can be obtained as follows:

$$\boldsymbol{F}_S = log(\mathcal{W}_M \cdot \|STFT(\boldsymbol{x})\|^2), \tag{1}$$

where $\mathcal{W}_M \in \mathbb{R}^{M\times B}$ represents the Mel-filter banks and B is the number of frequency bins of the spectrogram, obtained by short-time Fourier transform (STFT).

To compensate for the missing anomaly information from the log-Mel spectrogram, we apply a CNN-based network (TgramNet) to extract the temporal feature from the sound signal $\boldsymbol{x}$. The architecture of TgramNet is shown in Fig. 1(b) and Table 1. Firstly, a large kernel 1D convolution is used, with channel number, kernel size and stride set the same as the number of Mel bins, window size and hop length for the log-Mel spectrogram. Then, three CNN blocks are applied, and each block consists of a layer normalization [14], a leaky ReLU activation function, and a 1D convolution with a smaller kernel size. Note that, the CNN blocks do not change the dimension of the output temporal feature. Thus, the resultant temporal feature, Tgram, is

$$\boldsymbol{F}_T = TN(\boldsymbol{x}), \tag{2}$$

where $TN(\cdot)$ represents the TgramNet for feature extraction from time domain, and $\boldsymbol{F}_T \in \mathbb{R}^{M\times N}$ has the same dimension as $\boldsymbol{F}_S$.

Finally, the spectral-temporal fusion feature STgram $\boldsymbol{F}_{ST} \in \mathbb{R}^{2\times M\times N}$ is obtained using a simple fusion strategy

Table 2. Performance comparison in terms of AUC (%) and pAUC (%) for different types of machines.

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 [7]	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 [8]	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(CEE)	91.30	86.73	91.25	81.69	99.36	96.84	94.44	91.58	88.80	87.38	72.93	**63.62**	89.68	84.64
STgram-MFN(ArcFace)	**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**

by concatenating the log-Mel spectrogram $\boldsymbol{F}_S$ and Tgram $\boldsymbol{F}_T$, that

$$\boldsymbol{F}_{ST} = Concat(\boldsymbol{F}_S, \boldsymbol{F}_T), \tag{3}$$

where $Concat(\cdot)$ denotes the concatenation operation.

2.2. Self-supervised classification

We adopt a self-supervised classification strategy following [7], where the metadata (i.e., machine IDs) accompanying the audio feature (i.e., STgram) are used to learn feature representations of normal sound, resulting in the better ability of the model in distinguishing the normal and abnormal sound. Specifically, we choose MobileFaceNet (MFN) [13] as the baseline classifier to learn the delicate representation of normal sounds. The whole method is abbreviated as STgram-MFN. For better sensitivity to the anomalies, STgram-MFN applies ArcFace [15], rather than the cross-entropy error (CEE), as the loss function which helps increase the distance between classes and decrease the distance within classes.

3. EXPERIMENTS AND RESULTS

3.1. Experimental setup

Dataset We evaluate our method using DCASE 2020 challenge Task2 development and additional dataset [3], which consists of part of MIMII [16] and ToyADMOS dataset [17]. The MIMII dataset has four machine types (i.e., Fan, Pump, Slider and Valve), and each type includes seven different machines. The ToyADMOS dataset has two machine types (i.e., ToyCar and ToyConveyor), including seven and six different machines, respectively. Here, machine ID is used to identify different machines with the same machine type. In the experiments, the training data (normal sound) from the development and additional dataset of Task 2 [3] are used as the training set, and the test data (normal and anomaly sound) of the development dataset is employed for evaluation.

Evaluation metrics The performance is evaluated with the area under the receiver operating characteristic (ROC) curve (AUC) and the partial-AUC (pAUC), following [4,7,8], where pAUC is calculated as the AUC over a low false-positive-rate (FPR) range $[0, p]$ and $p = 0.1$ as in [3]. In addition, the minimum AUC (mAUC) is taken to represent the worst detection performance achieved among individual machines of the same machine type, following [8].

Implementation We train our proposed STgram-MFN on the training set of raw wave audio signals with a length of around 10 seconds, where one model is trained for all machine types. The frame size is 1024 samples with an overlapping 50% for the log-Mel spectrogram, and the number of Mel filter banks is 128 (i.e., $W = 1024$, $H = 512$ and $M = 128$). Accordingly, the obtained Sgram $\boldsymbol{F}_S$ and Tgram $\boldsymbol{F}_T$ have a dimension of 128×313. Adam optimizer [18] is employed for model training with a learning rate of 0.0001, and the cosine annealing strategy is adopted for learning rate decay. The model is trained with 200 epochs, and the batch size is 128. The margin and scale parameters of ArcFace [15] are 0.7 and 30, respectively. The negative log probability is used as the anomaly score for detection.

Table 3. Performance comparison in terms of mAUC (%).

Methods	IDNN [4]	Mobile NetV2 [7]	Glow_Aff [8]	**STgram -MFN (CEE)**	**STgram -MFN (ArcFace)**
Fan	56.56	50.40	49.60	79.80	**81.39**
Pump	61.86	52.90	65.70	79.79	**83.48**
Slider	74.22	82.80	87.80	**98.39**	98.22
Valve	66.83	67.90	77.70	79.12	**98.83**
ToyCar	64.41	55.70	80.10	61.91	**83.07**
ToyConveyor	62.89	48.70	61.00	57.25	**64.16**
Average	64.46	59.73	70.32	76.04	**84.86**

3.2. Performance comparison

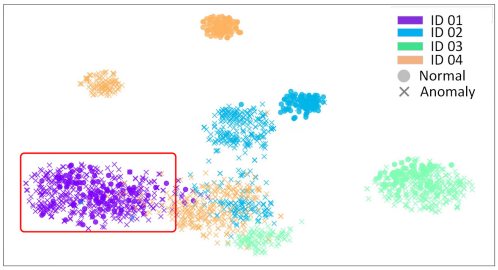
Table 2 shows the comparison of the proposed STgram-MFN with other state-of-the-art methods, IDNN [4], MobileNetV2 [7], and Glow_Aff [8]. IDNN is the AE-based method, MoblieNetV2 is the self-supervised classification method, and Glow_Aff is the flow-based self-supervised method. Regarding STgram-MFN, CEE and ArcFace loss are adopted for model training, respectively, denoted as STgram-MFN(CEE) and STgram-MFN(ArcFace). It is shown that the proposed method significantly improves the ASD performance, specifically 7.16% improvement on AUC and 8.6% improvement on pAUC (averaged over all the six machine types), compared with the best performance achieved by other methods in the literature.

Table 3 shows the mAUC results to reflect the worst detection performance achieved among individual machines of the same type. The instability of the previous methods can be observed that MobileNetV2 outperforms IDNN in Table 2, but its average mAUC (59.73%) is worse than IDNN (64.46%) in

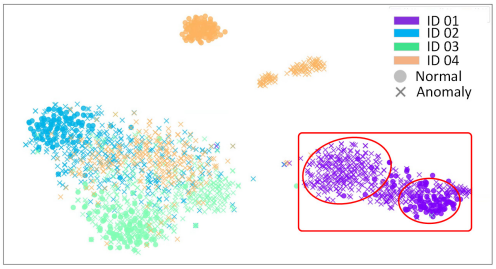
Table 4. Performance comparison for different input features.

Methods	LogMel-MFN		Tgram-MFN		Spec-MFN		**STgram-MFN(CEE)**		**STgram-MFN(ArcFace)**	
	AUC	mAUC	AUC	mAUC	AUC	mAUC	AUC	mAUC	AUC	mAUC
Fan	82.36	53.75	89.47	**83.85**	80.36	49.75	91.30	79.80	**94.04**	81.39
Pump	87.74	67.62	89.13	82.60	83.73	65.92	91.25	79.79	**91.94**	**83.48**
Slider	99.08	98.07	71.64	27.45	93.62	88.62	99.36	**98.39**	**99.55**	98.22
Valve	89.91	65.88	87.41	74.32	86.46	77.77	94.44	79.12	**99.64**	**98.83**
ToyCar	88.73	66.32	60.72	49.89	64.07	47.22	88.80	61.91	**94.44**	**83.07**
ToyConveyor	**78.17**	**67.79**	52.70	46.71	54.47	51.85	72.93	57.25	74.57	64.16
Average	87.67	69.91	75.18	60.80	77.12	63.52	89.68	76.04	**92.36**	**84.86**

Table 3. Amongst all the methods, STgram-MFNs achieves the best result with a greater mAUC improvement compared to Glow_Aff. Specifically, STgram-MFN (ArcFace) achieves the average mAUC of 84.86%, outperforming Glow_Aff by 14.54%.



(a) log-Mel spectrogram



(b) Tgram

Fig. 2. The t-SNE visualization of log-Mel spectrogram and Tgram for MFN classifier on the test dataset for the machine type Fan. Different color represents different machine ID. The "•" and "×" denote normal and anomalous classes, respectively. The normal and anomaly cluster for "ID 01" is highlighted by contours in red.

3.3. Ablation study

To show the effectiveness of STgram, we conducted an ablation study using log-Mel spectrogram, Tgram, spectrogram and STgram, respectively, as the input features of MFN, and the results are presented as LogMel-MFN, Tgram-MFN, Spec-MFN and STgram-MFN in Table 4. As ArcFace needs to adjust parameters for different methods, we only use CEE as the loss function in LogMel-MFN, Tgram-MFN and Spec-MFN for a fair comparison.

Table 4 shows that LogMel-MFN has a much smaller mAUC than AUC on Fan, Pump, Valve, and ToyCar, reflecting inconsistent performance for different machines even of the same type. Tgram-MFN performs better in terms of mAUC on Fan, Pump, and Valve than LogMel-MFN. However, log-Mel spectrogram and Tgram are complementary, as illustrated by t-distributed stochastic neighbor embedding (t-SNE) cluster visualization of the latent features of log-Mel spectrogram and Tgram in Fig. 2. For example, it is clear that the anomalous and normal latent features of machine "ID 01" are overlapping in terms of the log-Mel spectrogram, while they are more distinguishable by Tgram. This finding shows that the log-Mel spectrogram may filter out useful information about anomalies.

However, Tgram-MFN does not perform well in general, since it may suffer from noise contained in the temporal information. We also evaluated spectrogram without Mel filtering as the input feature (Spec-MFN in Table 4). It can be observed that Spec-MFN performs better on Slider and Valve but worse on Fan and Pump than Tgram-MFN. Compared to the above potential alternatives, the proposed input feature is much more effective for anomalous sound detection. STgram-MFN(CEE) and STgram-MFN(ArcFace) achieve much higher mAUC (76.04% and 84.86%, respectively), as compared with those of LogMel-MFN (69.91%), Tgram-MFN (60.80%), and Spec-MFN (63.52%), suggesting that they offer more stable detection performance.

4. CONCLUSION

In this paper, we have presented a self-supervised anomalous sound detection method, where a spectral-temporal fusion feature from the raw wave is applied, by combining temporal information from a CNN network and spectral information from the log-Mel spectrogram. The proposed method exploits complementary spectral-temporal information from the normal sound via the fused features, and results in more stable detection performance of amongst different machines. The experimental results demonstrated the effectiveness of the proposed method with substantial improvements over the state-of-the-art methods.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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 *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE)*, 2020, pp. 81–85.

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

[5] S. Kapka, "ID-conditioned auto-encoder for unsupervised anomaly detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE)*, 2020, pp. 71–75.

[6] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the Neyman-Pearson lemma," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 212–224, 2018.

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

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

[9] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *arXiv preprint arXiv:1807.03039*, 2018.

[10] G. Papamakarios, T. Pavlakou, and I. Murray, "Masked autoregressive flow for density estimation," *arXiv preprint arXiv:1705.07057*, 2017.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[12] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

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

[14] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

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

[16] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE)*, 2019, pp. 209–213.

[17] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, November 2019, pp. 308–312.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.