# UNSUPERVISED ANOMALOUS SOUND DETECTION FOR MACHINE CONDITION MONITORING USING TEMPORAL MODULATION FEATURES ON GAMMATONE AUDITORY FILTERBANK

Technical Report

*Kai Li, Quoc-Huy Nguyen, Yasuji Ota, Masashi Unoki*

Japan Advanced Institute of Science and Technology,
School of Information Science,
1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan
{kai_li, hqnguyen, y_ota, unoki}@jaist.ac.jp

## ABSTRACT

Anomalous sound detection (ASD) is a task to identify whether the sound emitted from a target machine is normal or not. Subjectively, timbral attributes, such as sharpness and roughness, are crucial for human beings to distinguish anomalous and normal sounds. However, the feature frequently used in existing methods for ASD is the log-mel-spectrogram, which cannot capture information in the time domain. This paper proposes an ASD method using temporal modulation features on the gammatone auditory filterbank (TMGF) to provide temporal characteristics for machine-learning-based methods. We evaluated the proposed method using the area under the ROC curve (AUC) and the partial area under the ROC curve (pAUC) with sounds recorded from seven kinds of machines. Compared with the baseline method of the DCASE2022 challenge, the proposed method provides a better ability for domain generalization, especially for machine sounds recorded from the valve.

***Index Terms—*** Anomalous sound detection, gammatone filterbank, temporal modulation features, timbre information, deep learning

## 1. INTRODUCTION

Anomalous sound detection (ASD) is a technique to judge whether the sound recorded from a target machine is normal or not. It allows workers to arrange maintenance work to fix machine problems in the earliest stages, thus reducing maintenance costs and preventing consequential damages. ASD for machine condition monitoring purposes has received increasing attention in recent years.

DCASE2022 Challenge Task2 [1] is a competition following the DCASE2020 Task2 [2] and DCASE2021 Task2 [3] for ASD using sounds emitted from different machines. There are mainly two challenges in this task. First, collecting anomalous sounds that can cover all possible types of anomalies is quite difficult. Therefore, this task is often viewed as an unsupervised problem. Second, the differences in acoustic characteristics between the training and test data, such as operational speed, machine load, and environmental noise shift, decrease the accuracy of the ASD system. Therefore, developing a method with a strong domain generalization ability is quite important.

DCASE2022 Task2 presented two baseline methods to deal with the above challenges. The first one is the autoencoder (AE)-based unsupervised method, which is a kind of inlier modeling (IM)-based detector. In the AE-based method, the anomaly score is calculated as the reconstruction error of the observed sound. To obtain small anomaly scores for normal sounds, the AE is trained to minimize the reconstruction error of the normal training data. This method is based on the assumption that the AE cannot reconstruct sounds that are not used in training, that is, unknown anomalous sounds. The second one is an outliner exposure (OE)-based detector that uses MobileNetV2. This baseline identifies from which section the observed signal was generated. In other words, it outputs the softmax value, which is the predicted probability for each section. The anomaly score is calculated as the averaged negative logit of the predicted probabilities for the correct section.

However, all baseline methods use the log-mel-spectrogram feature as input, which is difficult to capture information in the time domain. For human beings, it's pretty easy to distinguish anomalous and normal sounds by perceiving timbral attributes, such as sharpness and roughness. A feature that includes more timbral information is crucial.

We propose to use temporal modulation features on gammatone auditory filterbank (TMGF) in ASD task in DCASE2022 Task2 [1]. We assume that the TMGF feature can provide much more information related to human perception. The results show that our proposed method performs better in the target evaluation, which means the proposed method has a better ability for domain generalization.

## 2. BASELINE METHOD

The autoencoder (AE)-based system, which is often used as an unsupervised ASD system, was selected as baseline [1]. In the baseline system, the log-mel-spectrogram of the input audio $X = \{X_t\}_{t=1}^{T}$ was extracted and fed into an AE-based detector, where $X_t \in \mathbb{R}^F$, $F$ and $T$ are the number of mel-filters and time-frames, respectively. Then, the acoustic feature at $t$ is obtained by concatenating consecutive frames of the log-mel-spectrogram as $\psi_t = (X_t, ..., X_{t+P-1})$, where $D = P \times F$, $P$ is the number of frames of the context window. The anomaly score is calculated as:

$$A_\theta(X) = \frac{1}{DT} \sum_{t=1}^{T} ||\psi_t - \mathbb{F}(\psi_t)||_2^2, \qquad (1)$$

where $\mathbb{F}(\cdot)$ is the vector reconstructed by the AE model, and $||\cdot||_2$ is $\ell_2$ norm. The vector-reconstruct function $\mathbb{F}(\cdot)$ is simulated by the AE model. As shown in Fig. 1, The AE model includes

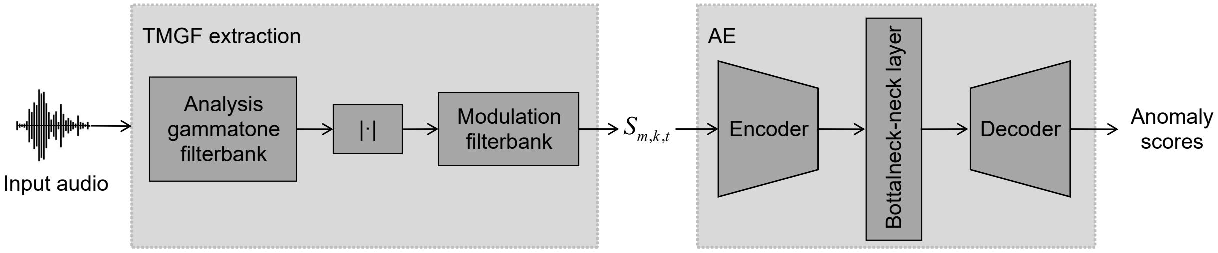

Figure 1: Proposed system using temporal modulation features on the gammatone auditory filterbank (TMGF, $S_{m,k}(t)$) for anomalous sound detection (ASD).

an encoder, bottleneck layer, and decoder modules. All modules consist of fully-connected layers. The training of the AE model is a regression mission due to only normal sound can be used in model training. Finally, the mean squared error (MSE) is used as the cost function to optimize the overall system.

To determine the anomaly detection threshold, the baseline method assumes that $A_\theta$ follows a gamma distribution. The gamma distribution parameters are estimated from the histogram of $A_\theta$, and the anomaly detection threshold is determined as the 90th percentile of the gamma distribution. If $A_\theta$ for each test clip is greater than this threshold, the clip is judged to be abnormal; if it is smaller, it is judged to be normal.

## 3. PROPOSED TMGF FEATURES

The temporal modulation on an auditory filterbank contains important information related to the timbre of a sound, such as the sharpness, roughness, and fluctuation strength [4, 5, 6]. Such information visualizes how humans perceive a sound as well as how we judge a sound (i.e., as ”anomalous” or ”normal”). Also, different frequencies of the temporal modulation contain different levels of speech information such as speech intelligibility, speaker identity, and emotion. Thus, we aim to utilize the temporal modulation feature for detecting anomalous sound. The extraction processes reference paper from Huy. et al. [7].

The gammatone filter [8] is a well-known auditory filter model. The impulse response of a gammatone analysis filter at the center frequency $f_c$ is defined as

$$g(t) = at^{n-1}e^{-2\pi b \mathrm{ERB}(f_c)t}e^{j2\pi f_c t}\,, \tag{2}$$

where $t \geq 0$ is time in seconds, $a$ is the amplitude, $n$ is the filter order, and $b$ is the bandwidth coefficient. The equivalent rectangular bandwidth $\mathrm{ERB}(f_c)$ is defined as

$$\mathrm{ERB}(f_c) = 24.7 + 0.108 f_c\,. \tag{3}$$

Using $K$ gammatone filters $\{g^{(k)}(t)\}_{k=0}^{K-1}$ with different center frequencies, from an input signal $x(t)$, the output of the filterbank $X_k(t)$ can be expressed as the product of the amplitude modulation $A_k(t)$ and the complex carrier $e^{j\phi_k(t)}$, as

$$\begin{aligned} X_k(t) &= x(t) * g^{(k)}(t) \\ &= A_k(t)e^{j\phi_k(t)}\,. \end{aligned} \tag{4}$$

The gammatone filterbank can be implemented using a wavelet transform where the mother wavelet is $\psi(t) = g(t)$ [9]. Then, with an $\alpha > 1$, the $k$-th filter $g^{(k)}(t)$ can be defined by scaling $\psi(t)$ with a factor $\alpha_k$ of $t$, as

$$g^{(k)}(t) = \psi\left(\alpha_k t\right)\,, \tag{5}$$

$$\alpha_k = \alpha^{\frac{2k}{K-1}-1}\,. \tag{6}$$

To analyze different frequency components of $A_{k,t}$, we use a modulation filterbank [10, 11] consisting of $M$ filters $\{h^{(m)}(t)\}_{m=1}^{M}$. The first filter $h^{(1)}(t)$ is a low-pass filter with the cut-off frequency of $f_1$. For each $m \geq 2$, the filter $h^{(m)}(t)$ is a band-pass filter of which the frequency ranges from $2^{m-2}f_1$ to $2^{m-1}f_1$. Using the designed modulation filterbank, the TMGF features can be extracted from the amplitude modulation $A_{k,t}$ as

$$S_{m,k}(t) = A_k(t) * h^{(m)}(t)\,. \tag{7}$$

## 4. EXPERIMENTAL SETUP

### 4.1. Datasets

The datasets used in this task were provided by the DCASE2022 organizers [12, 13]. The data includes the normal and anomalous sounds recorded from seven machines: fan, gearbox, bearing, slide, tor car, toy train, and valve. Each recorded sound includes the target machine’s sounds and environmental sounds. To simplify the task, only the first channel of multi-channel audio is used. The length of each recorded sound is fixed into 10s, and the sampling rate is 16 kHz.

The data is divided into three parts, development dataset, additional training dataset, and evaluation dataset. Each part includes audio from these seven types of machines. Machines in the development dataset include section 01, section 02, and section 03. Machines in the additional training dataset and evaluation dataset include section 04, section 05, and section 06. Each section was divided into source and target domains due to the differences in operating speed, machine load, viscosity, heating temperature, type of environmental noise, SNR, etc. Different domains are split into a training and testing subset—the training datasets include normal sounds only, but the testing datasets include normal and abnormal sounds.

In our experiments, training data in the development dataset was used for model training, and test data in the development dataset was used for testing. The submission results are based on the evaluation dataset.

Table 1: Overall results of the proposed (TMGF) method and baseline (BL) method in terms of AUC and pAUC.

| Machines | Sections | AUC (source) | | AUC (target) | | pAUC | |
|---|---|---|---|---|---|---|---|
| | | BL (%) | TMGF (%) | BL (%) | TMGF (%) | BL (%) | TMGF (%) |
| Toy car | 0 | 85.54 | 62.62 | 45.06 | 40.78 | 51.89 | 47.79 |
| | 1 | 87.22 | 67.66 | 42.02 | 39.76 | 53.53 | 48.42 |
| | 2 | 99.04 | 71.62 | 26.44 | **42.66** | 54.32 | **55.53** |
| | Arithmetic mean | 90.60 | 67.30 | 37.84 | **41.07** | 53.25 | 50.58 |
| | Harmonic mean | 90.22 | 67.10 | 35.79 | **41.03** | 53.23 | 50.35 |
| Toy train | 0 | 66.78 | 44.26 | 32.94 | 25.84 | 51.63 | 48.74 |
| | 1 | 77.56 | 61.82 | 30.58 | **45.92** | 50.37 | 49.37 |
| | 2 | 83.42 | 45.86 | 15.92 | **49.76** | 49.47 | **51.05** |
| | Arithmetic mean | 75.92 | 50.65 | 26.48 | **40.51** | 50.49 | 49.72 |
| | Harmonic mean | 75.27 | 49.53 | 23.83 | **37.23** | 50.48 | 49.70 |
| Bearing | 0 | 50.24 | 62.86 | 62.88 | **63.46** | 51.53 | **52.84** |
| | 1 | 66.12 | **66.44** | 63.96 | 62.42 | 52.79 | 49.53 |
| | 2 | 42.14 | **55.70** | 54.74 | **62.64** | 48.47 | **66.05** |
| | Arithmetic mean | 52.83 | **61.67** | 60.53 | **62.84** | 50.93 | **56.14** |
| | Harmonic mean | 51.06 | **61.33** | 60.23 | **62.84** | 50.86 | **55.29** |
| Fan | 0 | 82.04 | **84.20** | 38.66 | **42.00** | 59.63 | 50.11 |
| | 1 | 72.46 | 51.84 | 46.04 | **49.48** | 51.63 | 50.95 |
| | 2 | 81.84 | 78.58 | 65.64 | **67.50** | 63.89 | **64.37** |
| | Arithmetic mean | 78.78 | 71.54 | 50.11 | **52.99** | 58.39 | 55.14 |
| | Harmonic mean | 78.52 | 68.35 | 47.75 | **50.99** | 57.93 | 54.43 |
| Gearbox | 0 | 64.34 | 36.02 | 65.00 | 49.60 | 61.26 | 49.60 |
| | 1 | 65.84 | 59.22 | 57.40 | 54.86 | 53.63 | 50.58 |
| | 2 | 74.64 | 67.96 | 66.04 | **66.22** | 62.11 | 58.05 |
| | Arithmetic mean | 68.27 | 54.40 | 62.81 | 56.89 | 59.00 | 52.74 |
| | Harmonic mean | 67.98 | 50.54 | 62.57 | 56.08 | 58.74 | 52.48 |
| Slider | 0 | 80.42 | 46.26 | 56.82 | 45.12 | 62.21 | 48.26 |
| | 1 | 67.04 | 50.22 | 50.18 | **63.06** | 53.05 | 53.05 |
| | 2 | 86.78 | 23.88 | 40.82 | **53.60** | 54.37 | 48.37 |
| | Arithmetic mean | 78.08 | 40.12 | 49.27 | **53.93** | 56.54 | 49.89 |
| | Harmonic mean | 77.17 | 35.97 | 48.37 | **52.93** | 56.27 | 49.80 |
| Valve | 0 | 54.66 | **98.66** | 51.96 | **98.30** | 52.26 | **94.37** |
| | 1 | 50.58 | **59.80** | 52.06 | **60.94** | 49.95 | **54.16** |
| | 2 | 50.88 | **95.86** | 43.40 | **97.08** | 48.79 | **89.11** |
| | Arithmetic mean | 52.04 | **84.77** | 49.14 | **85.44** | 50.33 | **79.21** |
| | Harmonic mean | 51.98 | **80.45** | 48.78 | **81.34** | 50.29 | **74.47** |
| Average | Arithmetic mean | 70.93 | 61.49 | 48.03 | **56.24** | 54.13 | **56.20** |
| | Harmonic mean | 67.57 | 55.53 | 42.53 | **51.56** | 53.76 | **54.26** |

## 4.2. Metrics

To evaluate the performance of an ASD system, the area under the curve (AUC) and partial-AUC (pAUC) for receiver operating characteristic (ROC) curves are used. The pAUC is an AUC calculated from a portion of the ROC curve over the pre-specified range of interest. To increase the reliability, the pAUC is calculated as the AUC over a low false-positive-rate (FPR) range [0, p], where p=0.1 is used. The definition of AUC and pAUS for each machine type, section, and the domain is the same as the 2020 edition [2] and can be expressed as:

$$AUC_{m,n,d} = \frac{1}{N_d^- N_n^+} \sum_{i=1}^{N_d^-} \sum_{j=1}^{N_n^+} \mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \quad (8)$$

$$pAUC_{m,n} = \frac{1}{\lfloor pN_n^- \rfloor N_n^+} \sum_{i=1}^{\lfloor pN_n^- \rfloor} \sum_{j=1}^{N_n^+} \mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \quad (9)$$

where $m$ represents the index of a machine type, $n$ represents the index of a section, $d$ = source,target represents a domain, $\lfloor \cdot \rfloor$ is the flooring function, and $\mathcal{H}(x)$ returns 1 when $x > 0$ and 0 otherwise. $\{x_i^-\}_{i=1}^{N_-}$ and $\{x_j^+\}_{j=1}^{N_+}$ are normal and anomalous test clips in domain $d$ in section $n$ in machine type $m$, respectively. $N_-$ and $N_+$ are the number of normal and anomalous test clips in domain $d$ in section $n$ in machine type $m$, respectively.

## 4.3. Experimental conditions

To extract the log-mel-spectrogram feature, we first split 10s audios into different frames with frame lengths of 64ms and hop lengths of 32ms. Then, the Mel spectrogram is extracted using the

$melspectrogram$ module in the $librosa$ library with parameters as follows: n_fft=1024, hop_length=512, T=128 and power=2.0. Finally, five Mel-spectrogram features (P=5) were concatenated into one feature vector with a dimension of 640 and fed into the detector.

In the TMGF feature extraction, we used the gammatone filterbank with $K = 65$ and $\alpha = 10$. For the mother wavelet $\psi(t)$, we set $n = 4$, $b = 1.019$, and $f_c = 600\,\mathrm{Hz}$. For the modulation filterbank, we used $M = 6$ and $f_1 = 2\,\mathrm{Hz}$. To decrease the dimension of TMGF feature, dowmsampling was conducted to decrease the tempral dimension into 1600 Hz. Finally, feature vectors with a fixed dimension 390 were fed into the detector.

The model had four linear layers with 128 dimensions for the encoder, one bottle-neck layer with eight dimensions, and four linear layers with 128 dimensions for the decoder. The model was trained to minimize the error between the input feature vector $x$ and the reconstruction $x'$. We trained the model for 100 epochs using the Adam optimizer [14] with a learning rate of 0.0001 and a batch size of 128. The anomaly scores were calculated by the averaged reconstruction error.

## 5. RESULTS

The overall results were shown in Table 1. We compare the results using our proposed method with that of the baseline method. The improved results are highlighted in the table. From these results, we can find that the log-mel-spectrogram feature has been used to provide better performance in the source evaluations, but the performance significantly degrades in the target evaluation. On the other hand, the proposed method performs better in the target evaluation; even degradation occurs in the source evaluation, which means the TMGF feature can provide a better ability for domain generalization. Furthermore, results of the TMGF feature realize much better performance in both source and target evaluation in the valve. Finally, by using the TMGF feature, we improved the average arithmetic mean of AUC in the target evaluation from 48.03 % to 56.24 % and the average harmonic mean of AUC in the target evaluation from 42.53 % to 51.56 %.

## 6. CONCLUSION

This paper presented a method that combines the TMGF feature with an AE-based detector in the ASD challenges. With the proposed method, we aim to make up for the deficiency of the log-mel-spectrogram feature and provide more timbral information related to human auditory perception, such as sharpness and roughness. Experimental results in DCASE2022 Challenge Task2 showed that the proposed method could provide a better ability for domain generalization. Especially for machine sounds recorded from the valve, results from both source and target evaluation have significant improvements compared with the baseline method.

## 7. ACKNOWLEDGMENT

This work was SCOPE Program of Ministry of Internal Affairs and Communications (Grant Number: 201605002) and the Fund for the Promotion of Joint International Research (Fostering Joint International Research (B))(20KK0233).

## 8. REFERENCES

[1] 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 dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," *arXiv preprint arXiv:2206.05876*, 2022.

[2] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, *et al.*, "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2006.05822*, 2020.

[3] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on dcase 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," *arXiv preprint arXiv:2106.04492*, 2021.

[4] H. Fastl and E. Zwicker, *Psychoacoustics - Facts and Models*. Berlin: Springer, 2007.

[5] B. Moore, *An Introduction to the Psychology of Hearing (Sixth Edition)*. Brill, 2013.

[6] A. Pearce, T. Brookes, and R. Mason, "Timbral attributes for sound effect library searching," *Journal of The Audio Engineering Society*, pp. 2–2, 2017. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=18754

[7] Q.-H. Nguyen, K. Li, and M. Unoki, "Automatic mean opinion score estimation with temporal modulation features on gammatone filterbank for speech assessment," *Proc. IEEE-INTERSPEECH*, 2022.

[8] R. D. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," *Advances in speech, hearing and language processing*, vol. 3, pp. 547–563, 1996.

[9] M. Unoki and M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis," *Speech Communication*, vol. 27, pp. 261–279, 4 1999.

[10] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers," *The Journal of the Acoustical Society of America*, vol. 102, p. 2892, 6 1998. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.420344

[11] ——, "Modeling auditory processing of amplitude modulation. ii. spectral and temporal integration," *The Journal of the Acoustical Society of America*, vol. 102, p. 2906, 6 1998. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.420345

[12] 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 generalization task," *arXiv preprint arXiv:2205.13879*, 2022.

[13] 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," *arXiv preprint arXiv:2106.02369*, 2021.

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