Anomalous Sound Detection Based on Derivative Features of Short-Time Holomorphic Fourier Transform

Iori Hashimoto, Yu Morinaga, Suehiro Shimauchi, and Shigeaki Aoki
Kanazawa Institute of Technology, Japan
E-mail: c6401137@st.kanazawa-it.ac.jp

***Abstract*— Unsupervised anomalous sound detection (ASD) commonly employs an autoencoder (AE) with inputs typically consisting of short-time spectral features, such as short-time Fourier transform (STFT) spectrograms and mel-spectrograms. In this study, we focus on how the various input features used in unsupervised ASD differently affect detection performance, especially in a scenario where the derivative features of the short-time holomorphic Fourier transform (STHFT) are employed. The derivatives of the STHFT reveal the nonstationarities of both the amplitude and phase of short-time spectral components. The simulation results comparing the ASD performances with various input features indicated that the STHFT derivative features captured different signal characteristics in contrast to the conventional mel-spectrogram features.**

I. Introduction

Recently, anomalous sound detection (ASD), which evaluates machine functionality based on emitted sounds, has garnered attention amid the growing demand for automated condition monitoring and maintenance of industrial equipment, driven by workforce reductions and the widespread adoption of deep neural networks (DNNs).

DNN-based ASD methods can be broadly categorized into supervised and unsupervised learning approaches. Supervised approaches require a dataset that includes both "normal"-labeled and "anomalous"-labeled data to train the classifiers. However, in various scenarios, target machines rarely experience anomalies. Thus, it is practically difficult to collect sufficient "anomalous"-labeled data to train anomaly classifiers. This is a fundamental constraint in realizing ASD, in contrast to supervised learning-based applications such as sound event classification [1]. Therefore, semi-supervised approaches, which utilize a large amount of unlabeled data alongside a small amount of labeled data [2]–[6], and self-supervised approaches, which automatically generate labels from the unlabeled data, are often employed in practice [7]–[11]. However, even these supervised learning-based methods face challenges in handling unknown anomalies. In contrast, unsupervised approaches are more resistant to unknown anomalies, as they identify anomalies based on deviations from features learned from normal data. Unsupervised ASDs are trained exclusively on normal sound data, with autoencoders (AEs) commonly employed to detect anomalies based on reconstruction errors. Various AE architectures have been investigated to improve detection performance [12]–[18]. These AEs broadly employ short-time Fourier transform (STFT) spectrograms [12], [13] or mel-spectrograms [14]–[18] as inputs, though derivative features of short-time spectra, such as instantaneous frequency [19]–[21] and group delay [22], are also employed in other related fields.

In this study, we investigate the input features used for AEs in unsupervised ASDs. We employ derivative features of the short-time holomorphic Fourier transform (STHFT) [23], which are obtained by differentiating the STHFT for complex frequencies. The derivatives of the STHFT reveal the nonstationarities of both the amplitude and phase of short-time spectral components. We exploit nonstationary features by replacing conventional mel-spectrograms with STHFT derivative-based mel-spectrograms as inputs for AEs. This study aims to determine whether ASDs based on STHFT derivatives can capture distinct anomalous sound features compared with conventional ASDs based on mel-spectrograms. For training and evaluation, we use the dataset for DCASE 2024 Challenge Task 2 [24]–[26].

II. Anomalous Sound Detection System

Here, we consider the ASD system shown in Fig. 1. Input audio data are converted to a mel-spectrogram, which can be expressed in matrix form as,

$$X = \{X_m\}_{m=1}^{T}, \tag{1}$$

where, $X_m \in \mathbb{R}^F$, i.e., an F-dimensional column vector, and F and T are the number of mel-filters and time-frames, respectively. To input mel-spectrogram X sequentially into AE, small numbers of consecutive frames of X are concatenated as D-dimensional colomn vector, $\psi_m = (X'_m, \cdots, X'_{m+P-1})' \in \mathbb{R}^D$, where $D = P \times F$, P is the number of the sequential consecutive frames, and $'$ indicates transposition. Subsequently, the AE performs encoding and decoding ψ_m, reconstructing $\hat{\psi}_m$ as an approximation of the original ψ_m. By comparing the original ψ_m and the reconstructed $\hat{\psi}_m$, anomaly scores are

This work was partially supported by JSPS KAKENHI Grant Number JP22K12102.

979-8-3315-7206-8/25/$31.00

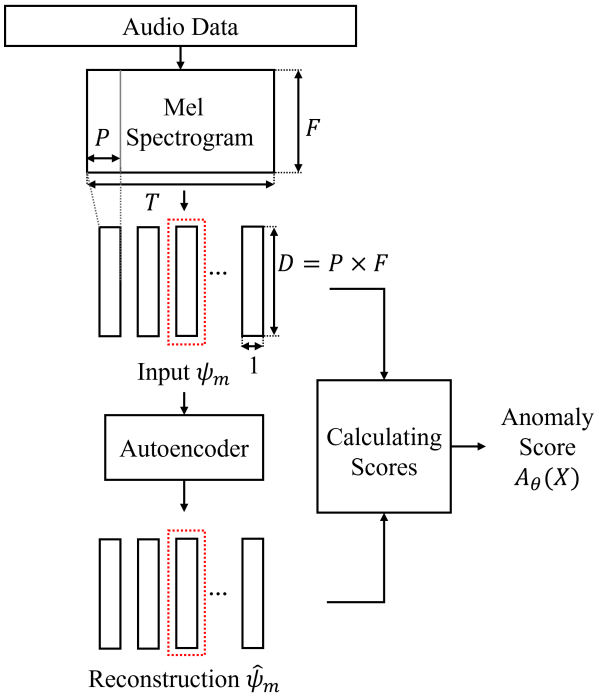

Fig. 1: Overview of ASD system.

calculated to detect anomalies. The details of each component are presented below.

A. Mel-Spectrogram Based on STFT Features

First, we describe the conventional mel-spectrogram based on STFT amplitude features. The STFT of audio signal $x(n)$ is obtained as follows:

$$X(\omega, m) = \sum_{n=-N}^{N-1} w(n)x(n+m)e^{-j\omega n}, \tag{2}$$

where n represents discrete time, m denotes time of interest, $w(n)$ is the window function, and ω denotes the discrete angular frequency. Each column vector component of the mel-spectrogram can be obtained by applying a triangular filter bank (mel-filter bank) $\mathcal{M}$ to the STFT amplitude $|X(\omega, m)|$ as follows:

$$X_m = \mathcal{M}(\{|X(\omega, m)|\}_{\omega=0}^{N}), \tag{3}$$

where the mel-filter bank $\mathcal{M}$ has F filters along with mel-scale optimized for human auditory properties [27]. The magnitude of the obtained mel-spectrogram is represented on a logarithmic scale (dB). Figure 2 presents an example of mel-spectrogram obtained from a normal sound of the 'gearbox' in the DCASE 2024 Challenge Task 2 dataset [26].

B. Mel-Spectrogram Based on STHFT Derivative Features

Here, we consider a mel-spectrogram based on the derivative features of the STHFT [23], which contrasts with the conventional mel-spectrogram based on the STFT amplitude. The

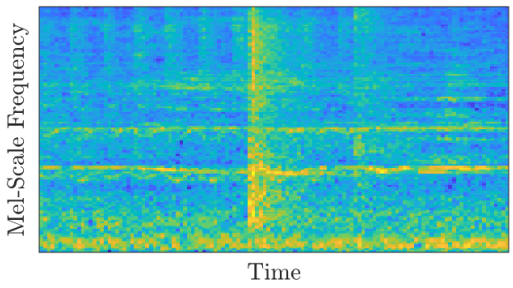

Fig. 2: Mel-spectrogram obtained from STFT amplitude.

definition of STHFT is as follows:

$$X(s, m) = \sum_{n=-N}^{N-1} w(n)x(n+m)e^{-sn}, \tag{4}$$

$$= A(s, m)e^{j\theta(s,m)}, \tag{5}$$

where $s = \alpha + j\omega$ is the complex frequency, α is introduced as an extension, $A(s, m)$ and $\theta(s, m)$ are the amplitude and phase of the STHFT $X(s, m)$. It is assumed that s remains continuous for differentiation but is normalized by the sampling frequency.

According to the discussion in [23], the partial derivative of logarithmic amplitude along with α, $\partial \log A(s, m)/\partial\alpha$, corresponds to the group delay, representing the increase or decrease of each spectral amplitude in the time direction. On the other hand, the partial derivative of the phase along with α, $\partial\theta(s, m)/\partial\alpha$, behaves like a relative instantaneous frequency, indicating the rate of amplitude change along the frequency axis at each time point. Based on these derivatives, the following weighted features are introduced, where the derivatives are weighted by the power spectrum $\mathcal{P}$ or $\mathcal{P}/(\mathcal{P}+1)$.

$$\mathrm{DF}_1(\omega, m) = \mathcal{P} \cdot \frac{\partial \log A(s, m)}{\partial\alpha}\bigg|_{s=j\omega}, \tag{6}$$

$$\mathrm{DF}_2(\omega, m) = \mathcal{P} \cdot \frac{\partial\theta(s, m)}{\partial\alpha}\bigg|_{s=j\omega}, \tag{7}$$

$$\mathrm{DF}_3(\omega, m) = \frac{\mathcal{P}}{\mathcal{P}+1} \cdot \frac{\partial \log A(s, m)}{\partial\alpha}\bigg|_{s=j\omega}, \tag{8}$$

$$\mathrm{DF}_4(\omega, m) = \frac{\mathcal{P}}{\mathcal{P}+1} \cdot \frac{\partial\theta(s, m)}{\partial\alpha}\bigg|_{s=j\omega}. \tag{9}$$

As in the previous section, a triangular filter bank is applied to these features to generate mel-spectrograms based on STHFT derivative features. Figures 3 and 4 show examples of mel-spectrograms obtained from (6) and (7), respectively, using the same signal analyzed in Fig. 2. Figure 3 highlights the temporal variations, whereas Fig. 4 highlights variations along frequency axis.

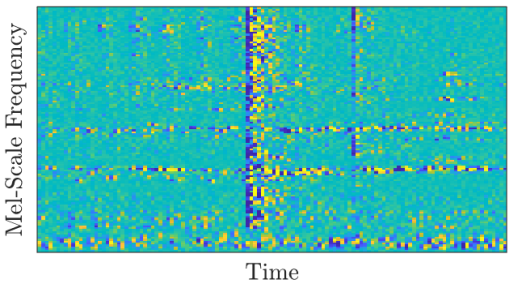

Fig. 3: Mel-spectrogram obtained from DF_1.

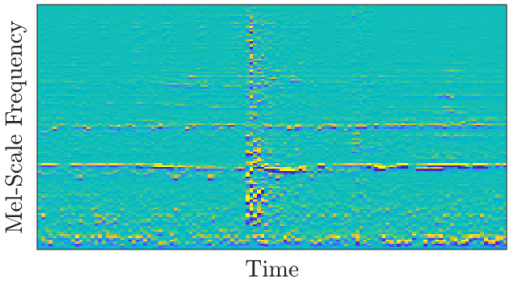

Fig. 4: Mel-spectrogram obtained from DF_2.

C. Autoencoder

An AE has a symmetrical structure consisting of an encoder and a decoder, as shown in Fig. 5. In this study, both the encoder and decoder are constructed as fully-connected DNNs with five layers each. The encoder extracts low-dimensional features common to the training data, while the decoder reconstructs the input signal from these features. As the encoder and decoder are trained to reconstruct the normal (non-anomalous) sound dataset by unsupervised learning, it is generally difficult to reconstruct anomalous sound inputs that are not used in training. Therefore, anomaly detection is facilitated by evaluating the reconstruction error between the AE input and output.

D. Anomaly Score Calculation and Detection

According to [28], an anomaly score that evaluates the anomalous degree of the input sound can be obtained as

$$\mathcal{A}_\theta(X) = \frac{1}{DT}\sum_{m=1}^{T}\|\psi_m - \hat{\psi}_m\|_2^2, \tag{10}$$

where $\mathcal{A}_\theta(X)$ is caluculated as the average of the mean-square error between the AE input ψ_m and the reconstructed output $\hat{\psi}_m$. Anomaly detection are conducted by judging whether the anomaly score $\mathcal{A}_\theta(X)$ exceeds a pre-defined threshold.

III. Simulations

We examined the influence of different types of mel-spectrograms as AE inputs on the performance of the ASD systems as described in the previous section. In particular, we compared ASD performance using a conventional STFT-based mel-spectrogram versus mel-spectrograms derived from each

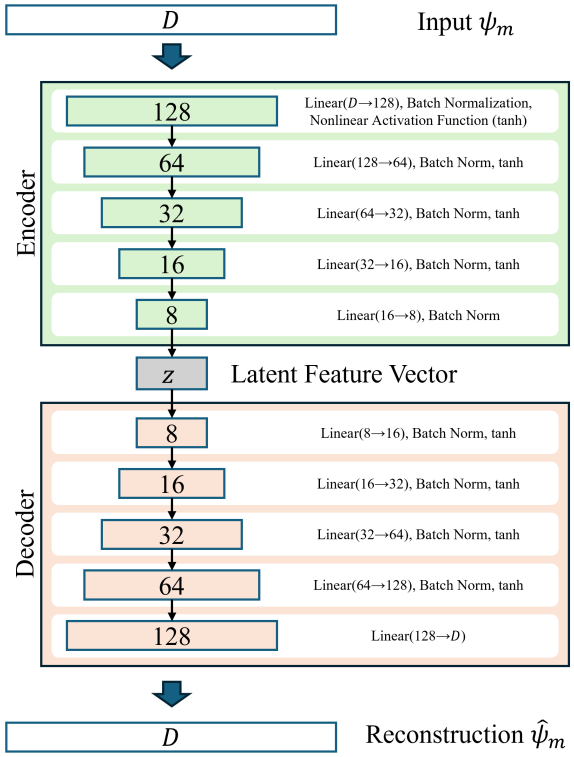

Fig. 5: Configuration of autoencoder.

of the STHFT derivative features defined in (6) to (9). Here, the ASD systems were implemented based on the DCASE 2024 Challenge Task 2 baseline system [17].

A. Dataset

The DCASE 2024 Challenge Task 2 development dataset [26] was used for training and evaluation. It consisted of normal/anomalous machine sounds of seven different machine types mixed with environmental noise, i.e., 'bearing', 'fan', 'gearbox', 'slider', 'ToyCar', 'ToyTrain', and 'valve'. For each machine type, dataset consisted of the audio clips recorded in two different domains, i.e., 'source' and 'target' domains, where some of the following conditions were different: operating speed, machine load, viscosity, heating temperature, environmental noise, and microphone arrangement, among others. The training dataset for each machine type consisted of 1000 clips: 990 normal clips from the 'source' domain and 10 normal clips from the 'target' domain, both without domain labels. The test dataset for each machine type included 200 clips, consisting of 50 normal and 50 anomalous clips from each domain (see Fig. 6). Each clip was a single-channel audio with a duration of 6–10 s.

B. Evaluation

The area under the receiver operating characteristic curve (AUC) and the partial AUC (pAUC) were used for the evalu-

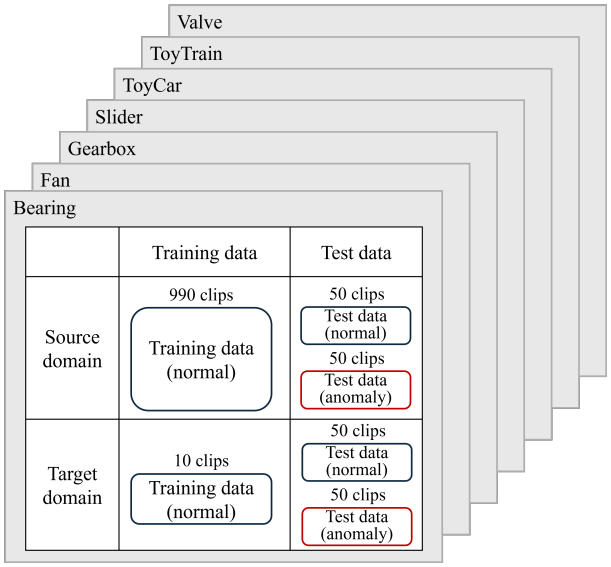

Fig. 6: Dataset [17].

ation, which are respectively defined as follows:

$$\text{AUC} = \frac{1}{I^- J^+} \sum_{i=1}^{I^-} \sum_{j=1}^{J^+} \mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \tag{11}$$

$$\text{pAUC} = \frac{1}{\lfloor pI^- \rfloor J^+} \sum_{i=1}^{\lfloor pI^- \rfloor J^+} \sum_{j=1}^{J^+} \mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)), \tag{12}$$

where I^- is the number of normal test clips and J^+ is the number of anomalous test clips, and $\mathcal{H}(x)$ returns 1 when $x > 0$ and 0 otherwise. The pAUC was calculated as AUC with a limited range $[0, p]$ of false positive rate (FPR), where $p = 0.1$ for our evaluations. According to the DCASE 2024 Challenge Task 2 evaluation guidelines, [28], the AUC was evaluated separately for each of data domains (source/target), using only normal test clips specific to each data domain, while the pAUC was evaluated as domain-independent. In addition, the harmonic mean of the AUC(source/target) and pAUC was evaluated as the 'Total Score'.

C. Parameter Settings

The frame size for the STFT and STHFT was 1024 (64 ms at 16 kHz sampling rate) and the hop size was 512. Hann window was applied to both the STFT and STHFT. The number of mel-filter banks was $F = 128$, the number of the sequential consecutive frames was $P = 3$, and thus, the length of one-dimensional input for AE was $D = P \times F = 384$. For training AEs, the batch size was set to 3732 and the number of epochs was set to 100. The learning rate was set to 0.001 for the STFT-based mel-spectrogram and 0.0001 for the STHFT derivative feature-based mel-spectrogram.

D. Comparison of Various Derivative Features of STHFT

First, we evaluated which of the STHFT derivative features, DF_1 to DF_4 defined in (6) to (9), were effective for the anomaly detection task, using a specific dataset 'bearing' from [26]. Table I shows the results, comparing with the STFT-based mel-spectrogram (referred to as BASE).

TABLE I: Performance comparison of different derivative features for 'bearing'.

	AUC (source)	AUC (target)	pAUC	Total Score
BASE	0.5758	**0.6626**	0.5637	0.5976
DF_1	0.5998	0.5454	0.5579	0.5668
DF_2	0.6192	0.5728	**0.6179**	0.6025
DF_3	0.6468	0.5920	0.5479	0.5928
DF_4	**0.6598**	0.6190	0.6158	**0.6309**

TABLE II: Performance comparison including anomaly score combinations.

	AUC (source)	AUC (target)	pAUC	Total Score
BASE	0.5758	0.6626	0.5637	0.5976
DF_2	0.6192	0.5728	**0.6179**	0.6025
DF_4	0.6598	0.6190	0.6158	0.6309
BASE & DF_2	0.6162	0.6240	0.5979	0.6125
BASE & DF_4	**0.6682**	**0.7102**	0.6011	**0.6567**

BASE achieved the highest score in AUC (target), while DF_2 achieved the highest in pAUC, and DF_4 achieved the highest in both AUC (source) and Total Score. However, DF_1 and DF_3 did not demonstrate significant performance across any evaluation metric. This is likely because both features, representing derivatives of logarithmic amplitude with respect to time, exhibited substantial frame-to-frame variation, making them unsuitable for dimensional compression by the AE. Therefore, among DF_1 to DF_4, DF_1 and DF_3 were excluded from the following experiments.

E. Combining Anomaly Scores

We explored the potential to enhance by combining anomaly scores derived from BASE system with those from either DF_2 or DF_4, selecting the lower of the two scores. Table II shows the results of the score combinations. As shown, the combination of BASE and DF_4 achieved the highest scores in AUC (source/target) and Total Score. Although selecting a lower score was a simple rule, certain enhancements in performance were observed.

TABLE III: Overall performance comparison.

	metric	bearing	fan	gearbox	slider	ToyCar	ToyTrain	valve
BASE	AUC(source)	0.5758	**0.7600**	0.5898	**0.6042**	**0.6284**	**0.6594**	**0.4724**
	AUC(target)	0.6626	0.4302	**0.6068**	0.5434	0.3754	0.4124	0.3728
	pAUC	0.5637	**0.5847**	0.5347	0.4963	0.4768	0.4795	0.5068
	Total Score	0.5976	0.5607	0.5754	0.5444	**0.4723**	0.4978	0.4430
DF_2	AUC(source)	0.6192	0.4126	0.4618	0.4756	0.4982	0.6398	0.4718
	AUC(target)	0.5728	**0.7408**	0.5302	0.5716	0.3644	**0.5750**	**0.5284**
	pAUC	**0.6179**	0.5268	0.5232	**0.5184**	0.5063	0.5142	0.4905
	Total Score	0.6025	0.5289	0.5031	0.5189	0.4460	**0.5718**	**0.4958**
DF_4	AUC(source)	0.6598	0.5054	0.6388	0.5204	0.5178	0.5952	0.4510
	AUC(target)	0.6190	0.6646	0.5490	0.4460	**0.3982**	0.5652	0.4766
	pAUC	0.6158	0.5695	0.5316	0.4916	**0.5084**	**0.5221**	0.4847
	Total Score	0.6309	**0.5726**	0.5695	0.4840	0.4681	0.5592	0.4703
BASE & DF_2	AUC(source)	0.6162	0.6090	0.5520	0.5428	0.5520	0.6324	0.4626
	AUC(target)	0.6240	0.5364	0.5718	**0.5980**	0.3556	0.4984	0.4700
	pAUC	0.5979	0.5142	**0.5726**	0.5026	0.4858	0.4889	**0.5184**
	Total Score	0.6125	0.5504	0.5653	**0.5451**	0.4490	0.5326	0.4824
BASE & DF_4	AUC(source)	**0.6682**	0.6868	**0.7018**	0.5600	0.5664	0.6138	0.4620
	AUC(target)	**0.7102**	0.4126	0.5472	0.5258	0.3754	0.5122	0.4298
	pAUC	0.6011	0.5405	0.5500	0.4926	0.4916	0.5021	0.5100
	Total Score	**0.6567**	0.5236	**0.5916**	0.5247	0.4641	0.5383	0.4650

F. Overall Evaluations

Finally, the performance was evaluated using each dataset of all machine types. The results are shown in Table III. Here, we focused on the Total Score as an essential metric for evaluating performance. In Table III, BASE achieved the highest Total Score only for 'ToyCar' dataset. DF_2 solely achieved the highest Total Score for 'ToyTrain' and 'valve' datasets. DF_4 solely achieved the highest Total Score for 'fan' dataset. Combinations of BASE and either DF_2 or DF_4 enhanced the Total Score performance for 'bearing', 'gearbox', and 'slider' datasets. These results indicate that ASDs based on STHFT derivatives-based mel-spectrograms DF_2 and DF_4 could capture distinct anomalous sound features compared with conventional ASDs based on STFT-based mel-spectrogram.

IV. Conclusions

In this study, we proposed ASD methods based on derivative features of STHFT. We investigated whether ASD based on STHFT derivative-based mel-spectrograms can capture different features from anomalous sound data compared to conventional mel-spectrogram-based ASDs, using a common DNN architecture. For the dataset and system configuration used in this study, the imaginary part of the STHFT derivative features demonstrated better performance. Based on the harmonic mean of AUC and pAUC, as a Total Score, ASDs utilizing STHFT derivative features, either alone or combined with conventional features, outperformed ASD based solely on conventional features across six of the seven machine types in the test dataset. The results suggest that the STHFT derivative features capture different signal characteristics compared to the conventional mel-spectrogram features.

References

[1] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust Sound Event Classification Using Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 540–552, 2015.

[2] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-Supervised Anomaly Detection via Adversarial Training," in *Asian conference on computer vision*, Springer, 2018, pp. 622–637.

[3] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, "TASK 2 DCASE 2020: Anomalous Sound Detection Using Unsupervised and Semi-Supervised Autoencoders and Gammtone Audio Representation," DCASE2020 Challenge, Tech. Rep., Jul. 2020.

[4] L. Ruff, R. A. Vandermeulen, N. Görnitz, *et al.*, *Deep Semi-Supervised Anomaly Detection*, Feb. 2020. arXiv: 1906.02694.

[5] X. Cai, H. Dinkel, Z. Yan, *et al.*, "A Contrastive Semi-Supervised Learning Framework for Anomaly Sound Detection.," in *DCASE*, 2021, pp. 31–34.

[6] L. Yang, J. Chen, Z. Wang, *et al.*, "Semi-Supervised Log-Based Anomaly Detection via Probabilistic Label Estimation," in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, IEEE, 2021, pp. 1448–1460.

[7] R. Giri, S. V. Tenneti, F. Cheng, K. Helwani, U. Isik, and A. Krishnaswamy, "Self-Supervised Classification for Detecting Anomalous Sounds," 2020. [Online]. Available: https://www.amazon.science/publications/self-supervised-classification-for-detecting-anomalous-sounds.

[8] K. Morita, T. Yano, and K. Tran, "Anomalous Sound Detection Using CNN-Based Features by Self Supervised Learning," *Tech. Rep., DCASE2021 Challenge*, 2021.

[9] K. Dohi, T. Endo, H. Purohit, R. Tanabe, and Y. Kawaguchi, "Flow-Based Self-Supervised Density Estimation for Anomalous Sound Detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 336–340.

[10] H. Hojjati and N. Armanfard, "Self-Supervised Acoustic Anomaly Detection via Contrastive Learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 3253–3257.

[11] K. Wilkinghoff, H. Yang, J. Ebbers, F. G. Germain, G. Wichern, and J. L. Roux, "Keeping the Balance: Anomaly Score Calculation for Domain Generalization," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India: IEEE, Apr. 2025, pp. 1–5.

[12] 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.

[13] J. Kim, Y. Jo, D. Lee, *et al.*, "Unsupervised Multi-View Reconstruction Autoencoder and Liquid Time Constant Model-Based First-Shot Anomaly Detection for Machine Condition Monitoring," DCASE2024 Challenge, Tech. Rep., Jun. 2024.

[14] A. Ribeiro, L. M. Matos, P. J. Pereira, *et al.*, *Deep Dense and Convolutional Autoencoders for Unsupervised Anomaly Detection in Machine Condition Sounds*, Jun. 2020. arXiv: 2006.10417.

[15] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, *Anomalous Sound Detection Based on Interpolation Deep Neural Network*, May 2020. arXiv: 2005.09234.

[16] E. Di Fiore, A. Ferraro, A. Galli, V. Moscato, and G. Sperlì, "An Anomalous Sound Detection Methodology for Predictive Maintenance," *Expert Systems with Applications*, vol. 209, p. 118 324, Dec. 2022.

[17] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-Shot Anomaly Sound Detection for Machine Condition Monitoring: A Domain Generalization Baseline," in *2023 31st European Signal Processing Conference (EUSIPCO)*, Helsinki, Finland: IEEE, Sep. 2023, pp. 191–195.

[18] J. Guan, Y. Liu, Q. Kong, *et al.*, "Transformer-Based Autoencoder with ID Constraint for Unsupervised Anomalous Sound Detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 42, Oct. 2023.

[19] A. M. Alqudah, "Towards Classifying Non-segmented Heart Sound Records Using Instantaneous Frequency Based Features," *Journal of Medical Engineering & Technology*, vol. 43, no. 7, pp. 418–430, 2019.

[20] C. C. Chatterjee, M. Mulimani, and S. G. Koolagudi, "Polyphonic Sound Event Detection Using Transposed Convolutional Recurrent Neural Network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 661–665.

[21] X. Chen and E. Kamavuako, "Enhancing Intake Monitoring: Transfer Learning for Audio-Based Detection of Swallowing Events," in *2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)*, 2024, pp. 479–484.

[22] A. Shah, P. Kevadiya, and H. A. Patil, "Pop Noise Detection Using Group Delay Cepstral Coefficients," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024, pp. 1–6.

[23] I. Hashimoto, Y. Morinaga, S. Shimauchi, and S. Aoki, "Derivative Features of Short-Time Holomorphic Fourier Transform," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 419–423.

[24] 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," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, Nov. 2021, pp. 1–5.

[25] K. Dohi, T. Nishida, H. Purohit, *et al.*, "MIMII DG: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection for Domain Generalization Task," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, Nov. 2022.

[26] T. Nishida, K. Imoto, N. Harada, *et al.*, *DCASE 2024 Challenge Task 2 Development Dataset*, Zenodo, Apr. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10902294.

[27] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "A Tutorial on Deep Learning for Music Information Retrieval," *arXiv preprint arXiv:1709.04396*, 2017.

[28] T. Nishida, N. Harada, D. Niizumi, *et al.*, *Description and Discussion on DCASE 2024 Challenge Task 2: First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring*, Jun. 2024. arXiv: 2406.07250.