Anomalous Sound Detection Using Time-Frequency Derivative of Instantaneous Phase Features

Tran-Quang-Tuan Vo*, Quoc-Huy Nguyen*, and Masashi Unoki*
* Japan Advanced Institute of Science and Technology
1–1 Asahidai, Nomi, Ishikawa 923–1292, Japan
E-mail: {tuan.vo, hqnguyen, unoki}@jaist.ac.jp

***Abstract*—Anomalous sound detection (ASD) plays a crucial role in predictive maintenance for industrial machines, enabling the monitoring of their health status through the analysis of sound signals. With the diversity of anomalous sounds in industrial environments and the scarcity of labeled data, unsupervised learning that leverages deep learning techniques demonstrates robust performance. However, the entire performance of the ASD detector depends significantly on the front end, which concentrates on feature extraction. While most ASD systems rely on amplitude information, there has been relatively little focus on phase information. This study proposes the use of the derivatives of instantaneous phase along time, frequency, and time-frequency, incorporating an Interpolation Deep Neural Network for ASD. The experimental procedures are conducted on the MIMII dataset, with the area under the receiver operating characteristic curve (ROC AUC) serving as the evaluation metric. The experimental results on the MIMII dataset demonstrate that our proposed method achieves an improvement in AUC over other recently proposed unsupervised methods that utilize amplitude information in detecting anomalous sounds from industrial machines.**

***Index Terms*—Anomalous sound detection, instantaneous phase time-frequency phase derivative**

I. Introduction

Anomalous sound detection (ASD) is the task of identifying whether the sound produced by a specific machine is normal or anomalous [1, 2]. Because anomalous sounds exhibit signs of malfunction, early detection and prevention can enhance predictive maintenance efforts, ultimately improving machinery reliability and reducing downtime. However, anomalous sounds originate from mechanical failures, which are characterized as zero-resource data due to their diverse nature. Most of the anomalous sound datasets used in this research field, such as MIMII [3] or ToyADMOS [4], are collected by deliberately damaging the target machines and are therefore impossible to simulate exhaustively. This reason highlights the importance of unsupervised anomaly detection systems, which can identify anomalies without requiring training on anomalous data.

Several techniques have been developed to tackle this issue. One such approach is a self-supervised method that employs pseudo-data to train the system [5–9]. While these techniques have shown promising comparative performance, they can be quite costly and heavily depend on data augmentation methods, which significantly impact the overall effectiveness of ASD systems. Another traditional approach involves inlier modeling with Autoencoder-based models [1, 10], which operates on the hypothesis that the acoustical characteristics of anomalous sounds cannot be accurately reconstructed by the model, as it is trained solely on normal data. Although these methods may appear straightforward, they rely heavily on the front end of the ASD system, also known as the feature extractor.

Acoustic features, such as log-Mel spectrograms, Short-Time Fourier Transform (STFT) spectrum [11], and timbral attributes [12], have proven effective and are widely used in various studies focused on the ASD task. Other studies have explored feature extraction using a Gammatone filterbank (GTFB) instead of a Mel filterbank [13–15], citing its superior representation of human auditory perception. Empirical results indicate that ASD systems utilizing Gammatone-based features outperform those using Mel-based features in detecting anomalous sounds. However, examining only the patterns of anomalous sounds based on the amplitude information of these features is insufficient for comprehensive detection. Sounds produced by mechanical failures often display unique patterns in their frequency or phase characteristics due to specific components of the machinery. By capturing these distinctive patterns in phase, the performance of the ASD system can be significantly enhanced.

Our previous research has explored the feasibility of using unwrapped instantaneous phase feature (IPGF) in ASD with an Autoencoder-based model [16], as extracted from the output of a GTFB. We hypothesized that interruptions in the phase trajectory indicate sudden changes in frequency due to abnormal events. The prior experimental results have suggested that distinguishable patterns exist between anomalous and normal sounds within the instantaneous phase information. If this hypothesis holds, detecting abnormal sounds could rely on interruptions in either the time or frequency axis; then, the fusion in both axes has the potential to enable a holistic detection. Those interruptions can be examined through the derivative of phase information.

This study proposes alternative representations of instantaneous phase features for ASD. This is achieved by calculating their derivatives along the time and/or frequency axes, and is referred to as the time derivative, frequency derivative, and time-frequency derivative of instantaneous phase features. To demonstrate the effectiveness of the proposed phase-based

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

features, we conduct experiments on the MIMII dataset using an Autoencoder Interpolation Deep Neural Network (IDNN) model as an anomaly detector. We compare the model's performance with similar models that utilize instantaneous amplitude (IAGF) [15] and IPGF features. Moreover, we also compare our obtained performance with that of other recently developed unsupervised methods.

This paper is structured as follows: In Section 2, we outline the mathematical derivation steps used to obtain three variants of phase representation. We also present a numerical approximation for implementing these derivatives. In Section 3, we describe our experimental scenario, focusing on the parameters, model configuration, and the dataset utilized. Finally, we present the experimental results and conclude with a summary of our study and direction for future research.

II. Proposed Features

The proposed features include three variants in phase representation: the time derivative, frequency derivative, and time-frequency derivative of instantaneous phase features. These features are derived from the Gammatone phase spectrogram.

A. Gammatone Phase Spectrogram

The Gammatone filterbank (GTFB) is a well-known auditory filterbank that simulates the response of the basilar membrane in the human auditory system [17]. The impulse response of the k^{th} filter with a center frequency f_k is expressed as

$$g(t) = At^{n-1}e^{-2\pi b\mathrm{ERB}(f_k)t}\cos 2\pi f_k t, \tag{1}$$

where $t \geq 0$ is the time in seconds, A, n, b are parameters, and $At^{n-1}e^{-2\pi b\mathrm{ERB}(f_k)t}$ is the amplitude term represented by the Gamma distribution of the k^{th} gammatone filter in the filterbank. The equivalent rectangular bandwidth (ERB) is defined as

$$\mathrm{ERB}(f_k) = 24.7 + 0.018 f_k. \tag{2}$$

To represent humans' auditory filter [18], we substitute the parameters in (1) with $n = 4$ and $b = 1.019$. Instantaneous information of an input signal $x(t)$ is obtained from the analytic representation of the filter, which is the Hilbert transform of (1) as

$$\psi(t) = At^{n-1}e^{j2\pi f_k t - 2\pi b\mathrm{ERB}(f_k)t}. \tag{3}$$

By using $\psi(t)$ to filter $x(t)$, we can obtain

$$X(k,t) = |X(k,t)|e^{j\theta(k,t)}, \tag{4}$$

with $|X(k,t)|$ is instantaneous amplitude spectrogram, while phase spectrogram is defined as

$$\theta(k,t) = \omega_k t + \phi(k,t), \tag{5}$$

with ω_k is the angular center frequency of the k^{th} filter in the filterbank, and $\phi(k,t)$ is the instantaneous phase in Radians and wrapped in its principal value, as $[-\pi, \pi)$.

The analytic Gammatone filterbank can be implemented using a bank of Finite-Impulse-Response (FIR) band-pass analytic Gammatone filters. The center frequencies are distributed linearly in the ERB scale where

$$r(f_k) = 21.4\log_{10}(0.00437 f_k + 1). \tag{6}$$

By substituting the ERB scale for the channel index in the notation of the filtered signal, the phase spectrogram and instantaneous phase can be articulated as a real multivariate function that relates center frequencies and time, as $\theta(r,t)$ and $\phi(r,t)$.

B. Time Derivative of Phase Feature

Time derivative of phase (TDP), also referred to as instantaneous frequency, indicates the rate of change of the instantaneous phase over time. TDP can be determined by calculating the first-order derivative of the unwrapped instantaneous phase in the continuous-time domain:

$$f(r,t) = \frac{1}{2\pi}\frac{\partial\widetilde{\phi}(r,t)}{\partial t}, \tag{7}$$

with $\widetilde{\phi}(r,t)$ in Radians is the unwrapped instantaneous phase in the temporal axis.

In the discrete-time domain, TDP can be numerically approximated by calculating the differences between adjacent values in the temporal axis. Leveraging the finite difference method, we can estimate the value of TDP as follows:

$$\frac{\partial\widetilde{\phi}(r,t)}{\partial t} = \widetilde{\phi}(r, t+\Delta t) - \widetilde{\phi}(r,t), \tag{8}$$

with Δt is the temporal distance between each sample in the temporal axis.

C. Frequency Derivative of Phase Feature

Frequency derivative of phase (FDP), also known as group delay, refers to the change in phase of a signal as a function of frequency. FDP can be calculated by determining the negative first-order derivative of the unwrapped phase spectrogram in the continuous-time domain:

$$\tau_g(r,t) = -\frac{\partial\widetilde{\theta}(r,t)}{\partial r}\frac{\partial r}{\partial\omega}, \tag{9}$$

with $\widetilde{\theta}(r,t)$ in Radians is the unwrapped phase spectrogram in the frequency axis.

In the discrete-time domain, the first term, $\frac{\partial\widetilde{\theta}(r,t)}{\partial r}$, is interpreted as the difference in phase between each channel, while the second term $\frac{\partial r}{\partial\omega}$ represents the differences in angular center frequencies among the channels. By utilizing the finite difference method, FDP can be approximated numerically as

$$\frac{\partial\widetilde{\theta}(r,t)}{\partial r}\frac{\partial r}{\partial\omega} = \frac{\widetilde{\theta}(r+\Delta r, t) - \widetilde{\theta}(r,t)}{\omega_c^{(r+\Delta r)} - \omega_c^{(r)}}, \tag{10}$$

with Δr is the distance between each center frequency of the filterbank in Cam units.

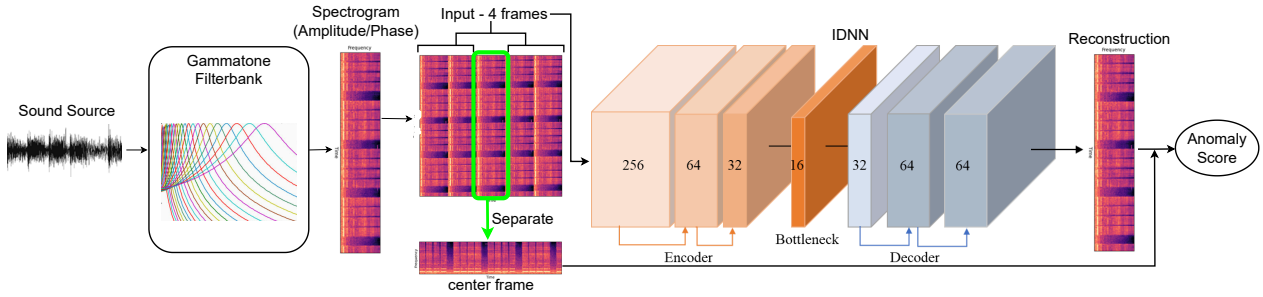

Fig. 1. Illustration of ASD based on spectrograms employing an IDNN model. The input sound is filtered with a GTFB to extract amplitude or phase-based features, which are then represented as spectrograms.

TABLE I
DATA DISTRIBUTION OF FOUR MACHINES IN THE MIMII DATASET

ID	Sound Types	Machine Types			
		Slider	Fan	Pump	Valve
00	Normal	1068	1011	1006	991
	Anomalous	356	407	143	119
02	Normal	1068	1016	1005	708
	Anomalous	267	359	111	120
04	Normal	534	1033	702	1000
	Anomalous	178	348	100	120
06	Normal	534	1015	1036	992
	Anomalous	89	361	102	120
Total	Normal	3204	4075	3749	3691
	Anomalous	890	1475	456	479

D. Time-frequency Derivative of Phase Feature

Time-frequency derivative of phase (TFDP) is defined as the second derivative of the instantaneous phase with respect to both the time and frequency axes. In this study, we first apply the temporal differentiation operation to the unwrapped instantaneous phase, which provides the instantaneous frequency information. The resulting data is then unwrapped and differentiated along the frequency axis. This procedure can be mathematically described as follows

$$\frac{\partial^2 \widetilde{\phi}(r,t)}{\partial\omega\partial t} = -\frac{\partial(\widetilde{2\pi f}(r,t))}{\partial r}\frac{\partial r}{\partial \omega}. \tag{11}$$

In the discrete-time domain, TFDP can be numerically approximated by employing the finite difference method, following the approach outlined in (8) and (10).

III. Experiment

A. Datasets

This study conducts experiments on the well-known MIMII dataset [3] with $\mathrm{SNR} = 6\,\mathrm{dB}$. The dataset comprises machine sounds recorded in a real factory environment, designed to examine and investigate machine faults through sound signal analysis. It includes both normal and abnormal sounds from four different types of machinery: Slider, Fan, Pump, and Valve. Due to the variations in mechanical components, the anomalous behaviors of each type are diverse and distinct. The systems are trained separately in an unsupervised manner using only normal data. Inference processes are conducted with both normal and abnormal data to evaluate overall performance. The statistics of normal and anomalous sound in MIMII are described in Table I.

B. Evaluation Metrics

This study utilizes the Area Under the Receiver Operating Characteristic curve (AUC-ROC) [19] as a comprehensive measure of a model's effectiveness in differentiating between the two classes, without relying on a specific anomaly threshold. AUC scores range from 0 to 1, with values approaching 1 indicating a greater likelihood that the model will accurately classify positive and negative sample pairs. This metric is computed as follows

$$\mathrm{AUC} = \frac{1}{\mathrm{N_-N_+}} \sum_{\mathrm{i}=1}^{\mathrm{N_-}} \sum_{\mathrm{j}=1}^{\mathrm{N_+}} \mathcal{H}(\mathcal{A}_\theta(\mathrm{x_j^+}) - \mathcal{A}_\theta(\mathrm{x_i^-})), \tag{12}$$

where $\{\mathrm{x_i^-}\}_{\mathrm{i}=1}^{\mathrm{N_-}}$ and $\{\mathrm{x_j^+}\}_{\mathrm{j}=1}^{\mathrm{N_+}}$ are the normal and anomalous test samples, with $\mathrm{N_-}$ normal samples and $\mathrm{N_+}$ abnormal samples. Additionally, $\mathcal{H}(x)$ returns 1 when $x = 0$ and otherwise. $\mathcal{A}_\theta(x)$ represents the anomaly score of sample sound x.

C. Autoencoder Interpolation Deep Neural Network

Traditional autoencoder-based models struggle to reconstruct the edge frames of concatenated spectrograms, particularly for non-stationary sound spectrograms. To address this issue, the Autoencoder Interpolation Deep Neural Network (IDNN) has been proposed [10]. This model focuses on interpolating only the missing center frame by effectively utilizing the information from the adjacent frames, specifically, the frames immediately to the left and right of the missing center frame.

Given an input $[x_1, \ldots, x_{\frac{n+1}{2}-1}, x_{\frac{n+1}{2}+1}, \ldots, x_n]$ frames and interpolate the frame $x_{\frac{n+1}{2}}$, the loss function of IDNN is expressed as

$$\mathcal{L}\left(x_{\frac{n+1}{2}} \middle| \mathcal{D}(\mathcal{E}([x_1, \ldots, x_{\frac{n+1}{2}-1}, x_{\frac{n+1}{2}+1}, \ldots, x_n]))\right), \tag{13}$$

where $\mathcal{E}, \mathcal{D}$ and $\mathcal{L}$ are the encoder, decoder, and loss function used in IDNN. In this work, we utilize IDNN as the backbone for anomalous sound detectors due to its effectiveness and superiority in reconstructing non-stationary spectrograms. The workflow of ASD using IDNN is depicted in Fig. 1.

TABLE II
SPECIFICATION OF UTILIZED IDNN MODEL

Components	Layer	No. of Units	Activation Function
Encoder	Input	256	ReLU
	Layer 1	64	ReLU
	Layer 2	32	ReLU
	Layer 3	16	ReLU
Decoder	Layer 4	32	ReLU
	Layer 5	64	Linear
	Output	64	None

D. Experimental Configuration

The features of TDP, FDP, and TFDP are extracted using a time-domain FIR GTFB. The center frequencies in the filterbank are distributed linearly on the ERB scale, ranging from 2 Cam to 32 Cam. Additionally, the resulting spectrograms are downsampled to reduce the temporal dimension by using a moving average linear rectangular window with a size of 400 samples and a hop size of 160 samples. The downsampled spectrograms are then concatenated from 5 frames to create 320-dimensional input vectors, which are subsequently fed into the model. All models are trained simultaneously for 200 epochs, with a batch size of 64, using the Adam optimizer [20] with a learning rate of 0.001, and mean squared error (MSE) to compute the reconstruction error. The specification of the utilized IDNN model is described in Table II.

IV. RESULTS

Table III presents the performance results of the IDNN model based on the AUC score, utilizing five different features. The names of the proposed features, including TDP, FDP, and TFDP, are highlighted in the table. Improvements compared to IAGF or IPGF are indicated in bold or underlined text. When a result is both highlighted and underlined, it signifies an improvement relative to both IAGF and IPGF.

The IDNN models that utilize TDP and TFDP features demonstrate superior effectiveness in detecting anomalous sounds from machinery, specifically Slider, Fan, and Pump, compared to the IAGF feature used in ASD. Notably, for the Pump, there is an approximate 5% increase in the AUC score when employing TDP or TFDP as discriminative features compared to IAGF. Additionally, the models leveraging TDP and TFDP features outperform both IAGF and IPGF in detecting abnormal sounds from the Fan. The results for the Fan also indicate that combining phase derivatives in both the time and frequency axes significantly enhances the detection of anomalous sounds, as reflected by the higher AUC score for TFDP compared to TDP or FDP. Furthermore, when it comes to bearing faults in the Slider, abnormal sounds can also be effectively detected using either TDP or TFDP features, with TDP showing a slight improvement over both IAGF and IPGF. The results also demonstrate significant performance improvements when utilizing either TDP or FDP to detect anomalies in sound from Valve machinery. However, these improvements do not match those achieved by models utilizing

TABLE III
PERFORMANCE COMPARISON IN AUC OF ASD EMPLOYING FIVE GAMMATONE FEATURES ACROSS DIFFERENT MACHINES, WITH PROPOSED FEATURES TDP, FDP, AND TFDP

Machine		Features				
		IAGF [15]	IPGF [16]	**TDP**	**FDP**	**TFDP**
Slider	ID 00	0.997	0.974	0.976	0.912	0.964
	ID 02	0.822	0.965	**0.954**	0.667	**0.960**
	ID 04	0.984	0.957	0.971	0.555	0.876
	ID 06	1.000	0.903	0.933	0.538	0.883
	Avg	0.951	0.950	**0.959**	0.668	0.921
Fan	ID 00	0.855	0.858	**0.893**	**0.959**	**0.949**
	ID 02	0.939	0.986	0.985	0.729	**0.991**
	ID 04	0.987	0.957	0.963	0.953	**0.992**
	ID 06	0.995	0.989	0.993	0.976	0.993
	Avg	0.944	0.948	**0.956**	0.904	**0.981**
Pump	ID 00	0.804	0.923	**0.910**	**0.825**	**0.904**
	ID 02	0.715	0.974	**0.953**	0.673	**0.897**
	ID 04	0.997	1.000	**1.000**	0.786	0.870
	ID 06	0.946	0.892	0.787	0.942	0.929
	Avg	0.866	0.947	**0.913**	0.807	**0.900**
Valve	ID 00	0.922	0.611	0.664	0.444	0.513
	ID 02	1.000	0.716	0.870	0.660	0.569
	ID 04	0.946	0.697	0.760	0.923	0.790
	ID 06	0.854	0.660	0.713	0.756	0.679
	Avg	0.931	0.660	0.752	0.696	0.638

IAGF features, which remain the most effective for identifying abnormal sounds from Valve. This poor performance may be attributed to the non-stationary and sparse nature of the valve's sound over time. Because the derivative of phase has an inverse relationship with amplitude information, the presence of silent segments in the valve sound can negatively impact the calculation of the reconstruction error between the target and the reconstructed phase spectrogram, ultimately compromising the effectiveness of the ASD models in detecting abnormal valve sounds when using phase-based features.

To illustrate the effectiveness of the proposed phase-based features, we employ t-SNE projection [21] to visualize the latent embedding of one of the machines in Fig. 2. Additionally, the original features are visualized to eliminate the model's effect and better understand the behavior of the proposed features. By observing the black contours, we can confirm that the proposed phase-based features effectively reflect the anomaly patterns of anomalous sounds, especially for the fan machine, whose abnormal behavior is caused by rotor-to-stator rubbing, resulting in phase interruption. Moreover, this visualization also demonstrates the effectiveness of fusing time and frequency derivatives of phase in detecting anomalous sound, as indicated by distinguishable clusters. Additionally, we acknowledge that the red contours highlight the slight challenges faced by the IDNN model in the structuring phase-

TABLE IV
COMPARISON RESULTS IN AUC OF THE PROPOSED METHOD AND OTHER UNSUPERVISED METHODS.

Machine	Methods							
	AE [1]	CVAE [22]	GRLNet [23]	IDNN+IAGF [15]	Deep SVDD [24]	**IDNN+TDP**	**IDNN+FDP**	**IDNN+TFDP**
Slider	0.902	0.894	0.911	0.951	0.867	**0.959**	0.668	0.921
Fan	0.953	0.903	0.953	0.944	0.936	0.956	0.904	**0.981**
Pump	0.868	0.887	0.901	0.866	0.837	**0.913**	0.807	0.900
Valve	0.594	0.655	0.639	**0.931**	0.804	0.752	0.696	0.638

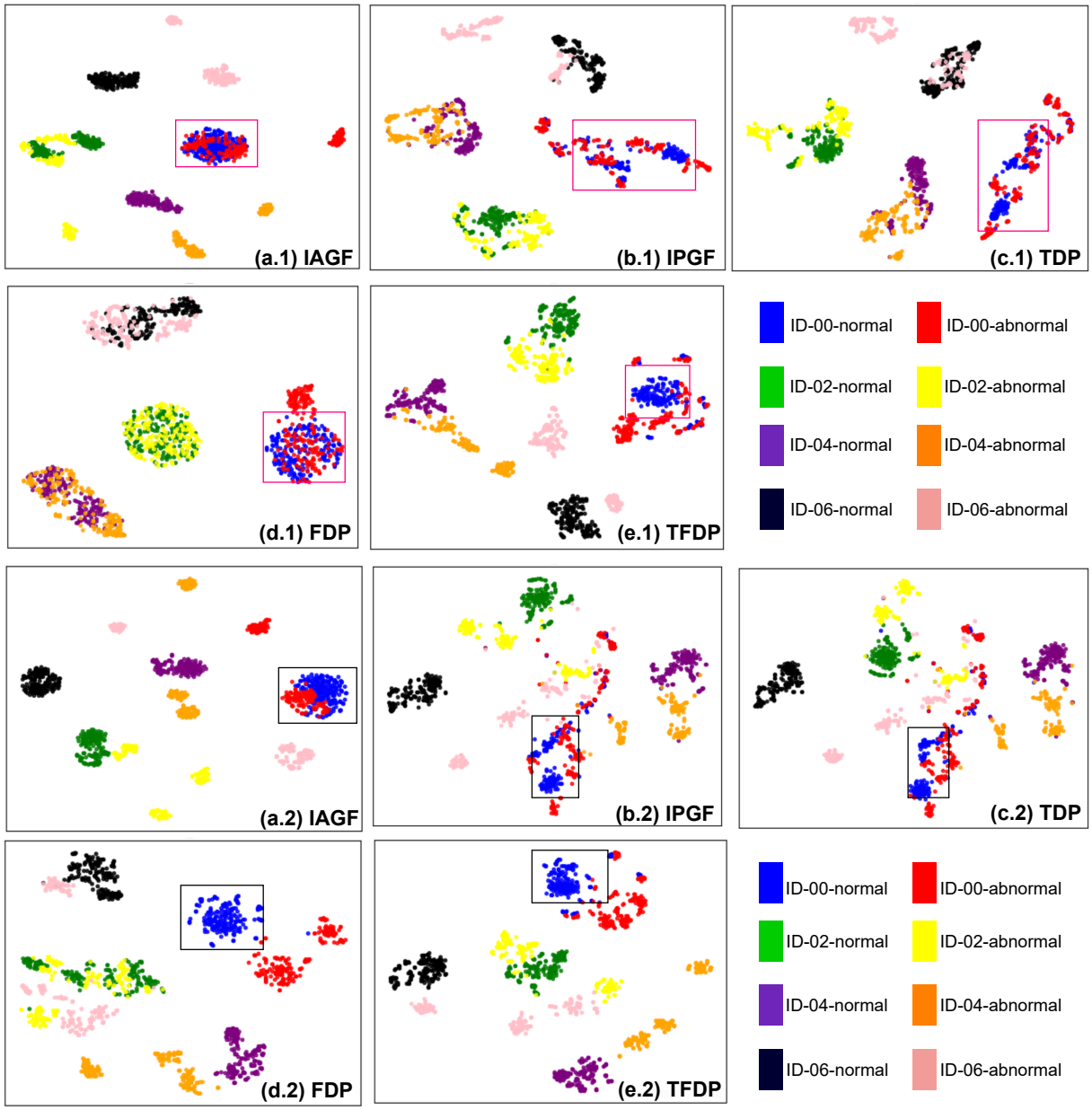

Fig. 2. The t-SNE visualization of IDNN bottleneck features (a.1)–(e.1) and original features (a.2)–(e.2) of IAGF, IPGF, TDP, FDP, and TFDP features of the Fan machine type in the MIMII dataset. Different colors represent different machine IDs and sound types. The colored contours demonstrate the significant discrimination ability of the proposed phase-based features in comparison with amplitude-based features.

based information to its latent space in ASD.

This study compares the performance of the proposed method, which utilizes IDNN and phase-based features, with that of recently developed unsupervised methods. These methods include DCASE 2020 Baseline [1], CVAE [22], GRLNet [23], IDNN+IAGF [15], and Deep SVDD [24]. These methods have used the amplitude information as the input feature. The performance comparison is described in Table IV. From these results, we observe that our proposed method outperforms others in detecting anomalous sounds from Slider, Fan, and Pump, and remains less competitive than the other methods in detecting abnormal sounds from Valve.

V. Conclusion

This study proposed three phase-based features, including the time derivative, frequency derivative, and time-frequency derivative of instantaneous phase features derived from the outputs of the Gammatone filterbank for unsupervised anomalous sound detection (ASD) using an Interpolation Deep Neural Network (IDNN) model. Our proposed method, which utilizes phase-based features, demonstrates an improvement over other recently proposed unsupervised methods that use amplitude information in detecting anomalous sounds from industrial machines such as Slider, Fan, and Pump in the MIMII dataset. However, the results indicated that our approach still encountered challenges in detecting abnormal valve sounds. The latent features from IDNN-based models exhibited some overlap between abnormal and normal clusters in t-SNE visualizations, highlighting a new challenge in this study: designing deep learning models optimized for phase-based features in the ASD task. Moreover, the future work should evaluate phase-based features for ASD under various Signal-to-noise ratio (SNR) conditions. Additionally, the fusion of both amplitude and phase information is also considered a potential strategy for creating a more comprehensive ASD system.

Acknowledgment

This work was supported by the JSPS Grant-in-Aid for Transformative Research Areas (A) (23H04344) and Grant-in-Aid for Challenging Research (Exploratory) (25K22817).

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 Task2: Unsupervised anomalous sound detection for machine condition monitoring," in Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, pp. 81–85, 2020.

[2] T. Nishida, T. Harada, N. Niizumi, D. Albertini, D. Sannino, R. Pradolini, S. Augusti, F. Imoto, K. Dohi, K. Purohit, H. Endo, T. and Kawaguchi, Y., "Description and discussion on DCASE 2024 Challenge Task2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, pp. 111–115, 2024.

[3] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa and Y. Kawaguchi, "MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection," in Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, pp. 209–213, 2019.

[4] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA), pp. 313–317, 2019.

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

[6] H. Hojjati and N. Armanfard, "Self-Supervised Acoustic Anomaly Detection Via Contrastive Learning," in Proc. ICASSP, pp. 3253–3257, 2022.

[7] H. Chen, Y. Song, L.-R. Dai, I. McLoughlin, and L. Liu, "Self-Supervised Representation Learning for Unsupervised Anomalous Sound Detection Under Domain Shift," in Proc. ICASSP, pp. 471–475, 2022.

[8] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in Proc. ICASSP, pp. 276–280, 2024.

[9] S. Ye and J. Yu, "SDMAE: A self-supervised learning method based on a self-distillation masked autoencoder for anomalous sound detection," Applied Acoustics, vol. 239, p. 110828, 2025.

[10] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in Proc. ICASSP, pp. 271–275, 2020.

[11] E. C. Nunes, "Anomalous sound detection with machine learning: A systematic review," arXiv preprint arXiv:2102.07820, 2021.

[12] Y. Ota and M. Unoki, "Anomalous Sound Detection for Industrial Machines Using Acoustical Features Related to Timbral Metrics," IEEE Access, vol. 11, pp. 70884–70897, 2023.

[13] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, and M. Cobos, "Anomalous sound detection using unsupervised and semi-supervised autoencoders and gammatone audio representation," arXiv preprint arXiv:2006.15321, 2020.

[14] K. Li, Q. H. Nguyen, Y. Ota, and M. Unoki, "Unsupervised anomalous sound detection for machine condition monitoring using temporal modulation features on gammatone auditory filterbank," in Proc. Detection Classification Acoustic Scenes Events (DCASE) Challenge, pp. 1–5, 2022.

[15] P. A. Hafiz, C. O. Mawalim, D. P. Lestari, S. Sakti and M. Unoki, "Anomalous Machine Sound Detection Based on Time Domain Gammatone Spectrogram Feature and IDNN Model," in Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Macau, Macao, pp. 1–6, 2024.

[16] Tran-Quang-Tuan Vo, Quoc-Huy Nguyen, and Masashi Unoki, "Feasibility of Anomalous Sound Detection by Utilizing Instantaneous Phase Features," in Proc. RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing 2025 (NCSP25), Pulau Pinang, Malaysia, pp. 57–60, 2025.

[17] R. D. Patterson, M. H. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," The Journal of the Acoustical Society of America, vol. 98, no. 4, pp. 1890–1894, 1995.

[18] 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, no. Part B, pp. 547–563, 1991.

[19] C. X. Ling, J. Huang, H. Zhang et al., "Auc: a statistically consistent and more discriminating measure than accuracy," in Ijcai, vol. 3, pp. 519–524, 2003.

[20] D.P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2015.

[21] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, vol. 9, no. 86, pp. 2579–2605, 2008.

[22] M.-H. Nguyen, D.-Q. Nguyen, D.-Q. Nguyen, C.-N. Pham, D. Bui, and H.-D. Han, "Deep convolutional variational autoencoder for anomalous sound detection," in Proc. IEEE 8th Int. Conf. Commun. Electron. (ICCE), pp. 313318, 2021.

[23] Y. Sha, S. Gou, J. Faber, B. Liu, W. Li, S. Schramm, H. Stoecker, T. Steckenreiter, D. Vnucec, N. Wetzstein et al., "Regional-local adversarially learned one-class classifier anomalous sound detection in global long-term space," in Proc. of ACM SIGKDD, pp. 3858–3868, 2022.

[24] S. Kilickaya et al., "Audio-based Anomaly Detection in Industrial Machines Using Deep One-Class Support Vector Data Description," IEEE Symposium on Computational Intelligence on Engineering/Cyber Physical Systems Companion (CIES Companion), pp. 1–5, 2025.