Master's Thesis

Study on anomalous sound detection using instantaneous phase features

VO Tran Quang Tuan

Supervisor UNOKI Masashi

Division of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

August, 2025

Abstract

Anomalous sound detection (ASD) is the task of identifying whether the sound produced by a specific machine is normal or anomalous. Because anomalous sounds exhibit signs of malfunction, early detection and prevention can enhance predictive maintenance efforts, ultimately improving machinery reliability and reducing downtime. Distinguishing between abnormal and normal sounds, a task that typically requires skilled and experienced machine engineers, faces significant challenges due to a shortage of human resources. The invention of ASD systems that leverage acoustical features related to human auditory perception is a promising solution for incorporating the strengths of both machine capabilities and human abilities to improve performance.

Most of the approaches in ASD concentrate on leveraging the superiority of deep-learning-based techniques, such as Autoencoder (AE) and acoustic features in the Mel scale, such as Mel spectrogram, etc., to model normal sound in an unsupervised manner in its latent space. The anomalous sound can be detected based on the large anomaly score after reconstructing the input spectrogram of AE-based models. Besides, the other approach, based on auditory perception analysis of Ota et al., attempts to tackle ASD by researching the primary differences between normal and anomalous sounds in hearing to develop timbral attributes. Despite obtaining attractive results in this approach, there exists a gap in ASD performance in detecting anomalous sound from some machine types in the MIMII dataset, which include Slider (ID 06) and Valve (ID 06). Based on the noticeable indicator of anomalous sounds emitted from these machine types, as argued by Ota et al., this study hypothesizes that the bearing faults of sliders or the beating sound of valves during malfunctions can cause sudden changes in the instantaneous frequency of these sounds. This hypothesis motivated this study to investigate the instantaneous phase and its derivative to detect phase interruptions, which represent the instantaneous changes in frequency better than amplitude caused by anomalous sound.

This study aims to propose a novel approach for ASD by utilizing instantaneous phase features. These features are derived from the outputs of an auditory filterbank, and then the derivative of phase is calculated along time, frequency, and time-frequency axes to capture interruptions holistically. The proposed phase-based features are presented in both the concepts of the derivation steps and the implementation. Later, the simulation with a frequency modulation signal is performed to validate the correctness of them. Moreover, a supervised experiment employing a support vector machine (SVM) and phase-based features is conducted on the MIMII dataset to verify the effectiveness in ASD.

Secondly, the unsupervised ASD system utilizing phase-based features is investigated in this study to handle the lack of anomalous data scenario. By leveraging the AE-based Interpolation Deep Neural Network (IDNN) model as the backbone and the Area Under the Receiver Operating Characteristic curve (AUC-ROC) as evaluation criteria, the experimental results demonstrate that the proposed phase-based features work well in detecting anomalous sound from most of the machine types in the MIMII dataset unsupervisedly, including Slider, Fan, and Pump, outperforming other unsupervised methods using amplitude-based features. Additionally, the study acknowledges the poor performance in detecting anomalous sounds from Valve. Therefore, further investigation of phase-based features in detecting anomalous valve sound is necessary.

In conclusion, this study achieved two research goals as presented, including proposing instantaneous phase features for ASD, verifying the correctness of the concepts, and establishing an unsupervised ASD system utilizing those. Through the experiment in an unsupervised manner, the proposed method demonstrates superior performance compared to other unsupervised methods using amplitude-based information as discriminated features. Future work should address the remaining drawbacks in this study, such as detecting anomalous sound under low SNR conditions, and improve the performance in detecting anomalous sound from Valve while using instantaneous phase features.

Contents

This thesis was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and VNU-HCMC University of Sciences.

List of Figures

List of Tables

List of Acronyms

AE	Autoencoder
APF	Amplified Predominant Frequency
AS	Amplified Shimmer
ASD	Anomalous Sound Detection
ERB	Equivalent Rectangular Bandwidth
FDP	Frequency Derivative of Phase
FIR	Finite Impulse Response
FM	Frequency Modulation
GMM	Gaussian Mixture Model
GTFB	Gammatone Filterbank
HMM	Hidden Markov Model
IA	Instantaneous Amplitude
IDNN	Interpolation Deep Neural Network
IM	Inlier Modeling
MCC	Mathews Correlation Coefficient
MFCC	Mel-frequency Cepstral Coefficients
OE	Outlier Exposure

SNR	Signal-to-noise Ratio
SVM	Support Vector Machine
TDP	Time Derivative of Phase
TFDP	Time-frequency Derivative of Phase
TF-ASD	Timbral-feature-based ASD
TMs	Timbral Metrics
UIP	Unwrapped Instantaneous Phase

Chapter 1

Introduction

1.1 Research background

Anomalous sound detection (ASD) is crucial for industrial companies to help workers and machine engineers arrange maintenance work, which reduces maintenance costs and prevents damage. The concrete indicator for failures or breakdowns is anomalous sounds emitted from industrial machines. Hearing the differences between abnormal and normal sound, which requires senior experience in machine engineering to discriminate between them, is facing many challenges due to a lack of human resources. ASD systems using acoustical features related to human auditory perception are a promising solution to deal with this problem. For example, damage points can be identified by using acoustical features used in ASD systems. Moreover, not only detecting but also predicting abnormal conditions by the ASD system helps prevent the interruption of operations in industrial equipment. Figure 1.1 illustrates the example of an ASD system.

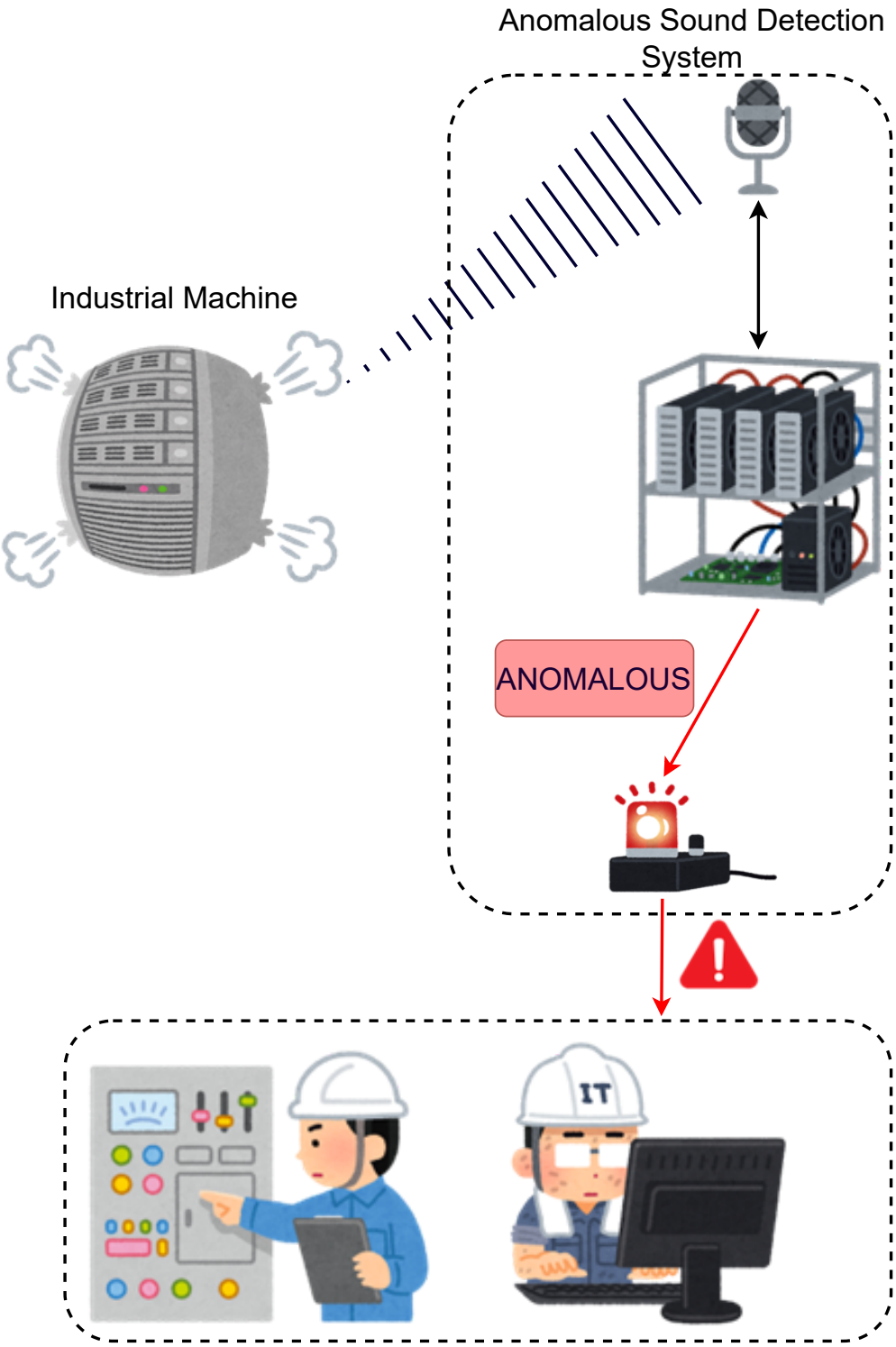

Figure 1.1: Illustration of an ASD system.

1.2 Problem statement

Anomalous sounds produced during machine operations indicate potential breakdowns in the early stages. Audio surveillance systems are crucial for monitoring machinery status [1]. Ota et al. [2] have proposed acoustic features related to timbral features for ASD in a supervised way and obtained attractive results. However, their experimental results on the MIMII dataset [3] at a signal-to-noise (SNR) ratio of 6 dB also reveal the poor performance of these features in detecting anomalous sound from some machine types, such as Slider (ID 06) and Valve (ID 06). This observation raises doubts about the ability of amplitude-based information to detect anomalous sound emitted from these machine types, while amplitude is the fundamental core of timbral features.

Based on the study of the mechanical architectures of industrial machines to investigate the characteristics of anomalous sound, Ota et al. have argued that most of the mechanical faults originate from the increase of friction between the components during malfunctions [2]. From the perceptual-based perspective, the friction can cause changes in the frequency of sounds from lower to higher suddenly, leading to a noticeable interruption in instantaneous frequency information. Therefore, the proposed method should be able to capture these artifacts in anomalous sound, thereby improving performance.

Additionally, the increasing variety of machine types often leads to abnormal behaviors that humans may struggle to detect, resulting in a limited amount of labeled data and the data imbalance problem for training anomaly detection systems in a supervised manner [4, 5]. With the rise of deep learning techniques, unsupervised learning presents a viable approach for anomaly detection in sound data [4, 6]. Therefore, the proposed method should facilitate the adaptation of deep-learning-based techniques to develop a robust unsupervised ASD system.

1.3 Research purpose

This study aims to propose an ASD method that utilizes instantaneous phase features related to auditory perception, incorporating deep learning tech-

niques, to achieve the research goals. In the proposed method, anomalous sound can be detected by capturing discontinuities/interruptions in instantaneous phase features that are caused by the anomalous sound. Additionally, to handle the scenario of the lack of anomalous data, the unsupervised ASD method utilizing instantaneous phase features will be investigated.

To achieve this research purpose, there are three primary tasks in this study. The first task is to develop the concepts of instantaneous phase features derived from the outputs of an auditory filterbank. Based on the hypothesis that phase interruption can exist in the acoustic characteristics of anomalous sound, the derivative of instantaneous phases along time, frequency, and both axes will be investigated. The second task involves verifying the proposed concepts through artificial simulation and validating the effectiveness of phase-based features in ASD by employing a supervised learning approach. The final task will concentrate on the second research goal, that is, integrating phase-based features with deep learning techniques to establish an unsupervised ASD system.

The originality of this study lies in its utilization of instantaneous phase information for ASD tasks. While the amplitude reflects variations in sound intensity levels, the phase holds valuable information about relative timing and frequency, and the instantaneous phase reflects the instantaneous variations of such information along the temporal axis. Although the instantaneous phase possesses a complex structure due to its cyclic wrapping nature [7,8], the discontinuity in its trajectory along the temporal axis can be observed due to a sudden change in the instantaneous frequency of anomalous sound. Therefore, the instantaneous phase can serve as an essential cue for effectively detecting sound anomalies in real-time applications. Moreover, the instantaneous phase features can be represented as spectrograms, which are particularly appropriate for integration with deep-learning approaches such as Autoencoder.

The significance of this study lies in both its application and scientific aspects. In the application aspect, this study contributes to the development of an ASD detection system across various industries, thereby enhancing safety, security, and quality by identifying anomalous sounds that may signal irregular events or malfunctions. From a scientific perspective, this study

contributes to the research in audio signal processing, utilizing instantaneous phase information, the potential of which has yet to be fully explored.

1.4 Structure of the thesis

This thesis is organized according to the following structure:

- Chapter 1 presents the significance of anomalous sound detection in Section 1.1, explains the problems in Section 1.2, and then describes and highlights the objectives, originality, and importance of this research in Section 1.3. Finally, Section 1.4 provides an overview of the thesis structure.

- Chapter 2 reviews the existing literature on anomalous sound detection. Section 2.1 discusses related DCASE Challenges, while Sections 2.2, 2.3, and 2.4 focus on the three main approaches in this research area: conventional, deep-learning-based, and auditory-perception-based approaches. Section 2.5 discusses the research issues.

- Chapter 3 focuses on developing concepts related to phase-based features that are derived directly from the outputs of an auditory filterbank, as described in Section 3.1. Section 3.2 outlines the validation process for these proposed phase-based features. Finally, Section 3.3 presents a supervised learning approach used to assess the effectiveness of these phase-based features in detecting anomalous sounds.

- Chapter 4 outlines the unsupervised approach for anomalous sound detection utilizing phase-based features in Section 4.1. Additionally, the practical implementation of this system is presented in Section 4.2.

- Chapter 5 presents the total evaluation for the unsupervised anomalous sound detection system utilizing phase-based features, as discussed in Chapter 4. Moreover, the comparative performance evaluation with other unsupervised methods is also conducted in this Chapter.

- Chapter 6 summarizes this thesis story, highlights the contribution, and discusses the prospects for future research.

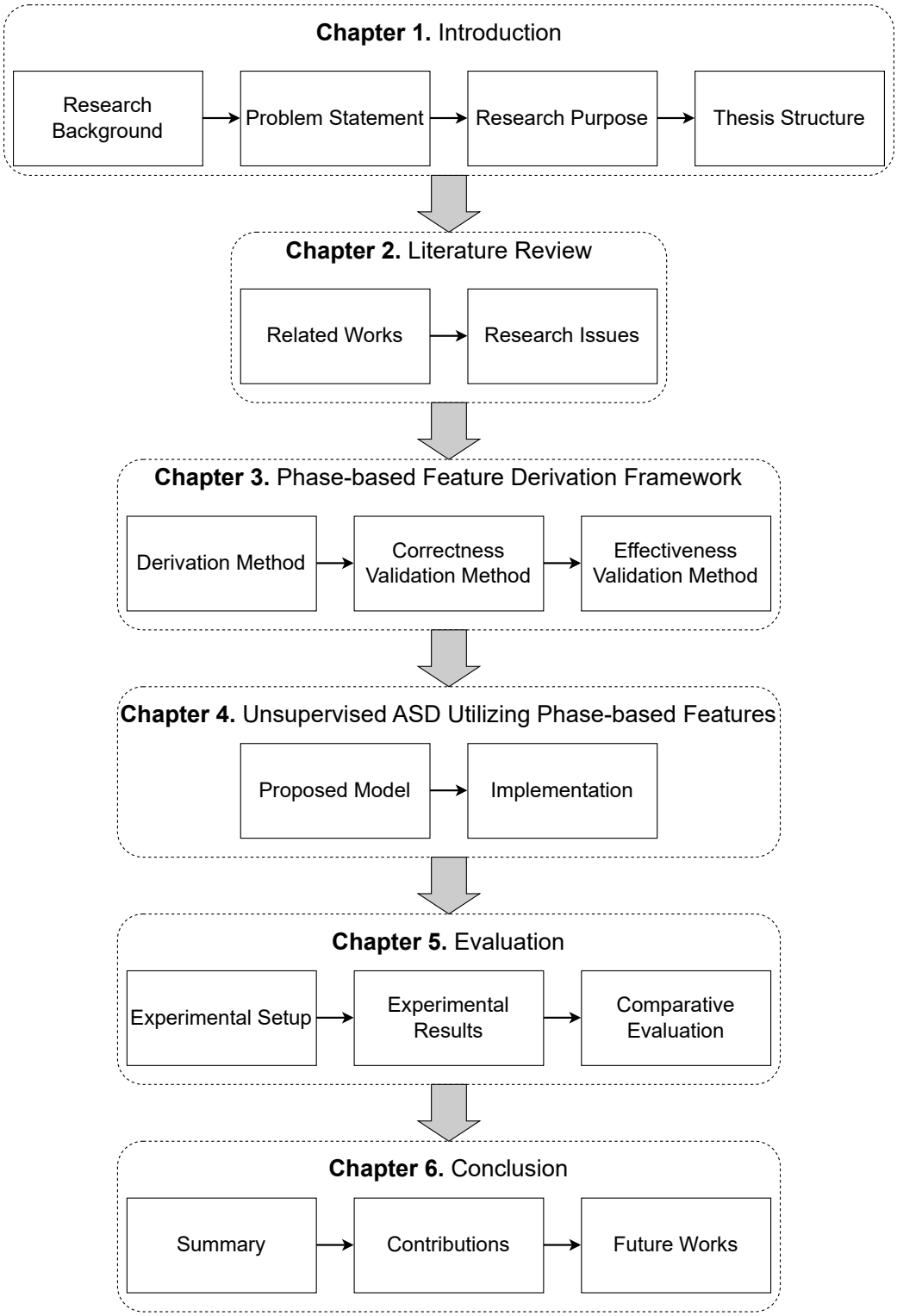

Figure 1.2: Structure of the thesis.

Chapter 2

Literature review

2.1 DCASE Challenge Task 2

The DCASE Challenge Task 2 is held annually to encourage research on detecting anomalous sounds from industrial machines. As real-world conditions evolve, the challenge's requirements have become increasingly complex, necessitating the integration of advanced and powerful processing techniques. The diagram in Figure 2.1 depicts the primary requirements of this challenge series. Below is a brief overview of them from 2020 to the present:

- The DCASE Challenge Task 2 - **Unsupervised Detection of Anomalous Sounds for Machine Condition Monitoring** in 2020 [5] marked the beginning of this series of challenges. Recognizing that real-world anomalous sounds are often more diverse than simulated ones, this challenge required systems to detect anomalous sounds in an unsupervised manner, relying solely on the normal sounds provided in the dataset.

- The normal sounds produced by machines can vary under different conditions, such as changes in weather, environmental factors, or the materials used in the engine. When a detection system is trained on the training dataset, it may struggle to identify these sounds accurately if the conditions in the test dataset diverge from those in the training dataset. This shift in the normal operating conditions can lead to misidentifications, making it difficult for the system to correctly detect

anomalous sounds. This motivated the organization of DCASE Challenge Task 2 - **Unsupervised Anomalous Sound Detection for Machine Condition Monitoring under Domain Shifted Conditions** in 2021 [9], which required systems to effectively detect anomalous sounds under domain shift conditions.

- One potential approach to addressing the domain shift problem is to use domain adaptation techniques. However, this method can be costly in real-world applications, which makes domain generalization a more preferred solution. The primary goal of domain generalization is to train the system using data from the source domain while enabling it to generalize the essential characteristics of this data. This allows the system to detect anomalous sounds in both the source and target domains. In light of these requirements, the DCASE Challenge Task 2 - **Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques** was organized in 2022 [10].

- Anomalous sound detection systems in previous years' challenges could only operate on sounds from machines that had been previously trained on. This issue presents a real-world scenario: the system's hyperparameters cannot be easily adjusted to adapt to entirely new machines. Therefore, developing a system that can be trained without the need for manual hyperparameter tuning is essential for ASD. Additionally, since some types of machines are limited in availability, the system must be capable of training with only a few machines from each type. These requirements opened a new topic in DCASE Challenge Task 2 2023 - **First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring** [11]. Furthermore, hiding the attribute information of machines in the dataset also adds to the difficulty of the challenge in 2024 [12].

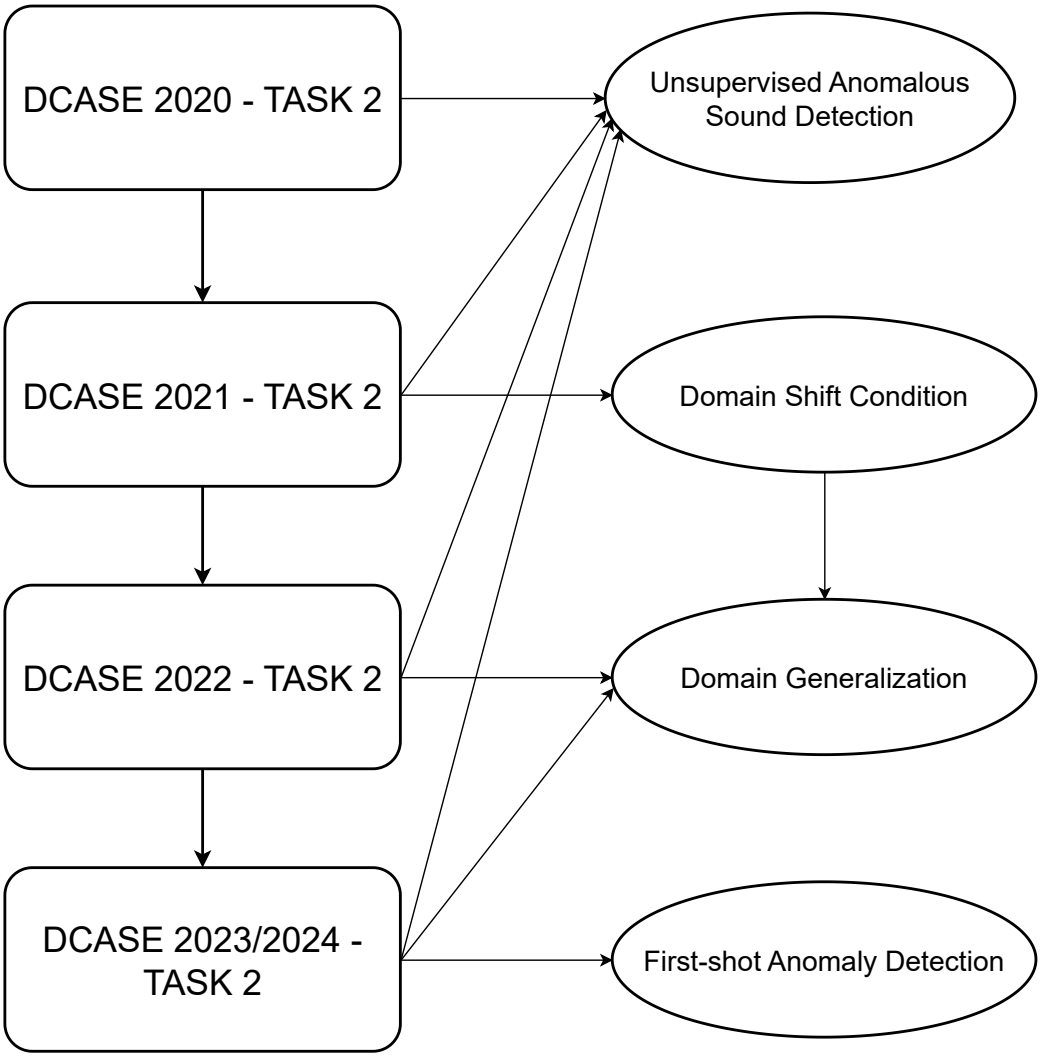

Figure 2.1: Requirements of DCASE Challenge Task 2

Various datasets have been proposed and utilized in the DCASE challenges for ASD. Among these, the two most prominent are the MIMII dataset [3] and the ToyADMOS dataset [13]. Additionally, other datasets such as MIMII DUE [14], MIMII DG [16], and ToyADMOS2 [15] have been introduced to address tasks related to domain shift and domain generalization. In the traditional unsupervised ASD task of the DCASE Challenge 2020, participants were given a development dataset containing sound samples from seven different types of machinery: Slider, Fan, Pump, and Valve (all sourced from the MIMII dataset), as well as ToyCar and ToyConveyor (from the ToyADMOS dataset). Furthermore, the development dataset was mixed with various types of environmental noise to better simulate real-world conditions [5].

2.2 Conventional approach

The conventional approach in ASD primarily employs lightweight machine learning techniques, utilizing acoustic features for discrimination. Ito et al. [17] have proposed an unsupervised ASD for surveillance microphones by leveraging the Multi-stage Gaussian Mixture Model (GMM) to build a model of normal sound, and the anomalous sound can be detected based on the likelihood threshold. Chan et al. [18] have investigated the temporal, spectral, parametric, and harmonic features and Hidden Markov Models (HMM) for scream classification. Another method has leveraged clustering techniques with Kullback-Leibler divergence to measure the dissimilarity between two distributions and detect anomalies [19]. Despite obtaining reasonable results under some conditions, these approaches are only suitable for small-scale datasets. Moreover, the traditional machine learning approach is less comparable in performance and scalability to the deep-learning-based approach.

2.3 Deep-learning-based approach

Most studies in this approach aim to enhance the ASD system in two key aspects: feature representation techniques and anomaly detection using deep

learning techniques.

Feature representation techniques

First and foremost, hand-crafted acoustic features have proven to be effective in distinguishing anomalous sounds. By capturing distinctive patterns, these acoustic features yield reliable results for ASD systems. Most ASD systems employ the Mel filterbank to extract the time-frequency characteristics of sound, which are then represented as spectrograms, including log-Mel spectrograms and log-Mel energies, to facilitate integration with deep learning techniques [5, 20]. Additionally, Hoang et al. [21] argue that the log-Mel spectrogram overlooks certain temporal characteristics. To address this, they propose a mixed feature approach that combines five types of acoustic features: Mel-frequency cepstral coefficients (MFCC), Chroma features, Mel spectrogram, Spectral Contrast, and Tonnetz, which serve as input for a U-Net-based detector. While this combination has shown performance improvements, the specific contributions of each feature in detecting anomalous sounds remain unclear. Other studies have explored feature extraction using a Gammatone filterbank instead of a Mel filterbank [22–24], 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.

Most ASD systems typically rely on hand-crafted acoustic features as their baseline. However, these features are affected by environmental noise, particularly in industrial settings. Moreover, hand-crafted acoustic features may not effectively capture hidden information related to anomalous sounds when compared to learnable features derived from deep learning models. As a result, several studies have explored the use of learnable features obtained from advanced deep learning techniques. For instance, Hayashi et al. [25] proposed an ASD method that employs WaveNet to model a variety of acoustic patterns in the time domain, allowing the system to identify unfamiliar acoustic patterns as anomalous sounds. Similarly, Han et al. [26] utilized pre-trained models such as Wave2Vec 2.0 [27], UniSpeech [28], and HuBERT [29] for ASD. Their results highlight the effectiveness of this approach in the DCASE 2023 Task 2, which addresses challenges related to

domain generalization and first-shot conditions in ASD.

Anomaly detection based on deep learning techniques
Anomaly detection for ASD based on deep learning techniques can be categorized into two primary approaches: supervised and unsupervised approaches [4, 6]. While supervised learning leverages both anomalous data and normal data for training the system, the unsupervised learning only needs to be trained with normal data to better capture the characteristics of normal sound. Then the anomaly can be detected if the learned features are far from the anomalies. Due to the scarcity of anomalous data, the unsupervised approach is widely applicable. In unsupervised ASD, Autoencoder (AE) based models are often used as the baseline [5]. The AE-based approach is a type of inlier modeling (IM), which is trained solely on normal data and determines anomalies based on the reconstruction error [30]. Another approach, which operates unsupervised, is outlier exposure (OE) [31], which utilizes external data as pseudo-anomalies to enhance detection ability. In addition, self-supervised approaches are also applied for ASD [32–35] to handle various complex scenarios given by DCASE. However, this approach is out of the scope of this thesis.

2.4 Auditory-perception-based approach

In addition to deep learning-based methods, other approaches investigate the primary characteristics of anomalous sounds through auditory perception analysis. A typical method in this category is the approach of Ota et al. [2] in proposing timbral attributes for ASD.

2.4.1 Timbral attributes for ASD

Based on investigating the key to discriminate anomalous sound from the perspective of timbre, Ota et al. have proposed a timbral-feature-based ASD (TF-ASD) using timbral metrics (TMs), including sharpness, roughness, boominess, brightness, and depth. Additionally, two short-term features, including amplified shimmer (AS) and amplified predominant fre-

quency (APF), are also proposed. The combination of these features is represented as the input for a support vector machine (SVM) classifier, with training in a supervised manner.

That research also describes the noticeable difference in hearing anomalous sounds emitted from industrial machines, based on the timbre perspectives. By using onomatopoeia, the study has explained the anomalous sound in the MIMII dataset clearly, such as the squealing sound from bearing faults of sliders, booming and whizz sounds from fans, splashing sound from pumps, and clicking and beating sound from valves, etc. It links this evidence with the timbral attributes to better understand the anomalous behaviors.

Through the experiment conducted on the MIMII dataset at a signal-to-noise ratio (SNR) of 6 dB and evaluated with accuracy, F1-score, and Matthew's Correlation Coefficient (MCC) metrics, TF-ASD using TMs+AS+APF demonstrated high performance in detecting anomalous sound, with average MCC for Slider, Fan, Pump, and Valve being 0.927, 0.976, 0.938, and 0.740, respectively.

2.5 Research issues

While the approach using timbral attributes has shown effectiveness in detecting anomalous sounds from industrial machines in the MIMII dataset, there are two primary issues that need to be addressed:

(1) Performance gap in detecting anomalous sounds from some machine types.
Although the previous work has reported positive results using timbral attributes for ASD, challenges remain in detecting anomalous sounds from certain machine types, specifically the Slider (ID 06) and Valve (ID 06). The low MCC and F1-score in the performance evaluation of the supervised ASD system for these machines highlight these difficulties.

(2) Supervised ASD is ineffective in the scenario of insufficient anomalous data.
The previous work has leveraged an SVM classifier and trained a classification

system supervisedly. Actually, 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 [13], 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.

Those two issues should be addressed in this study, contributing to the research of ASD based on the auditory-perception-based approach and improving the performance of ASD systems. Based on the hypothesis regarding phase interruption in anomalous sound, which was discussed in Section 1.2, an ASD method that utilizes instantaneous phase features will be proposed to tackle the first issue. Later, an unsupervised ASD method that uses phase-based features will be investigated to address the second issue.

Chapter 3

Phase-based features derivation framework

This chapter presents the framework for deriving and validating phase-based features to address the first research issue. The input signal is analyzed using an analytic auditory filter bank, which generates an analytic representation from which the phase spectrogram is derived. Additionally, a frequency-modulated signal with artificial phase interruptions is simulated to validate the derivation process. Furthermore, a supervised classification is conducted using the MIMII dataset to determine whether phase-based features can effectively detect anomalies in industrial machine sounds.

3.1 Gammatone phase-based features

The Gammatone filterbank (GTFB) is a well-known auditory filterbank that simulates the response of the basilar membrane in the human auditory system [36]. The impulse response of the k^{th} filter with a center frequency f_k is expressed as

$$g(t) = At^{n-1}e^{-2\pi b\mathrm{ERB}(f_k)t}\cos 2\pi f_k t, \tag{3.1}$$

where $t \geqslant 0$ 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

$$\mathrm{ERB}(f_k) = 24.7 + 0.018 f_k. \tag{3.2}$$

To represent humans' auditory filter [37], the parameters in Eq. (3.1) are substituted 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 (3.1) as

$$\psi(t) = A t^{n-1} e^{j2\pi f_k t - 2\pi b \mathrm{ERB}(f_k) t}. \tag{3.3}$$

The impulse response and frequency response of an analytic Gammatone filter are illustrated in Figure 3.1.

By using $\psi(t)$ to filter $x(t)$, the analytic signal is obtained:

$$X(k,t) = |X(k,t)| e^{j\theta(k,t)}, \tag{3.4}$$

with $X(k,t)$ is a bank of k sub-band signals, $|X(k,t)|$ is the instantaneous amplitude spectrogram, while the phase spectrogram is defined as

$$\theta(k,t) = \omega_k t + \phi(k,t), \tag{3.5}$$

with ω_k is the angular center frequency of the k^{th} filter in filterbank, and $\phi(k,t)$ is the instantaneous phase spectrogram in Radians. Additionally, the phase and instantaneous phase information of each sub-band signal are 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 of filterbank are distributed linearly in the ERB scale, where

$$r(f_k) = 21.4 \log_{10}(0.00437 f_k + 1). \tag{3.6}$$

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 $\phi(r,t)$.

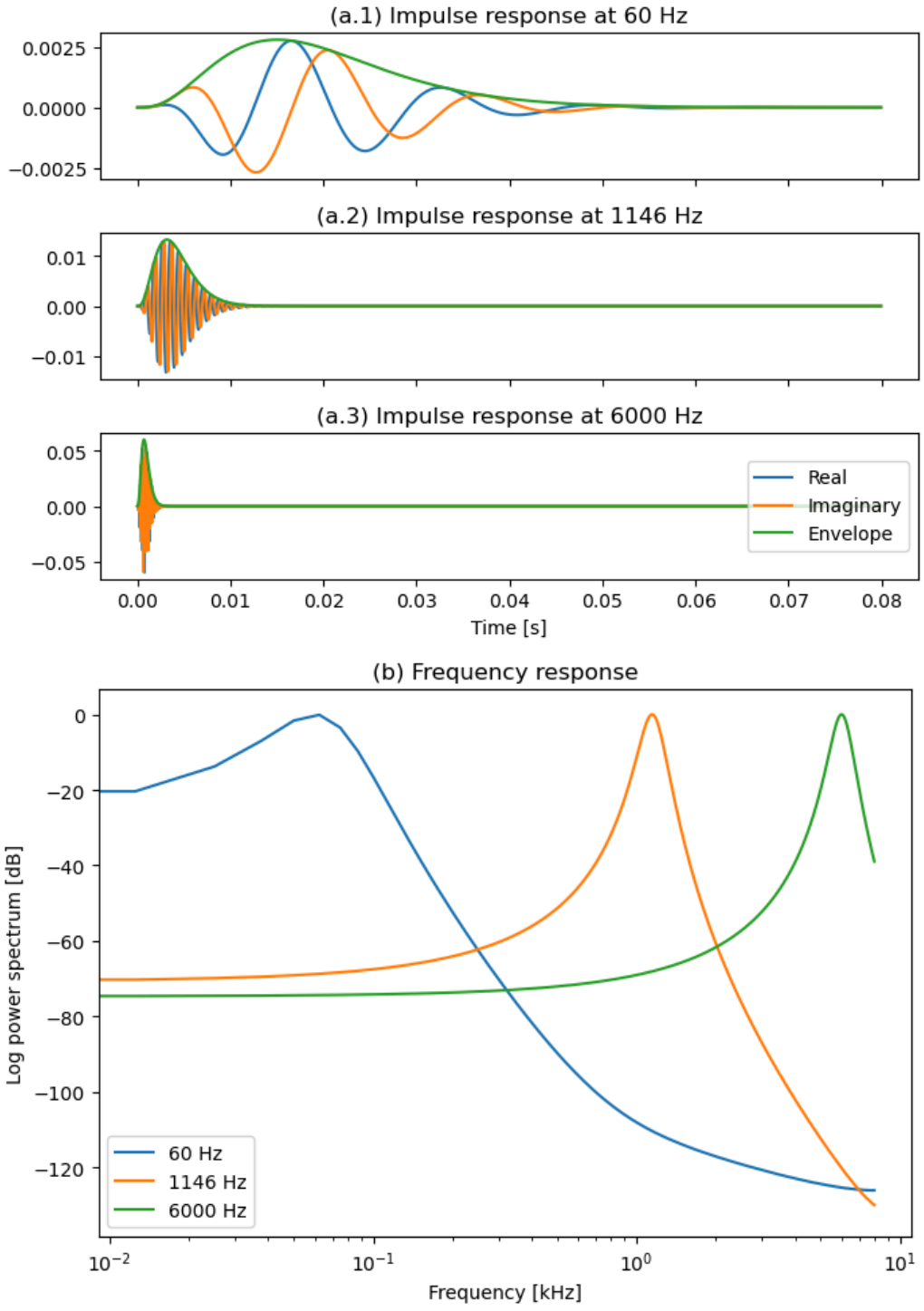

Figure 3.1: Illustration of impulse responses (a.1)-(a.3) and frequency responses (b) of Gammatone filters at center frequencies 60 Hz, 1146 Hz, and 6000 Hz of an analytic GTFB with 64 channels.

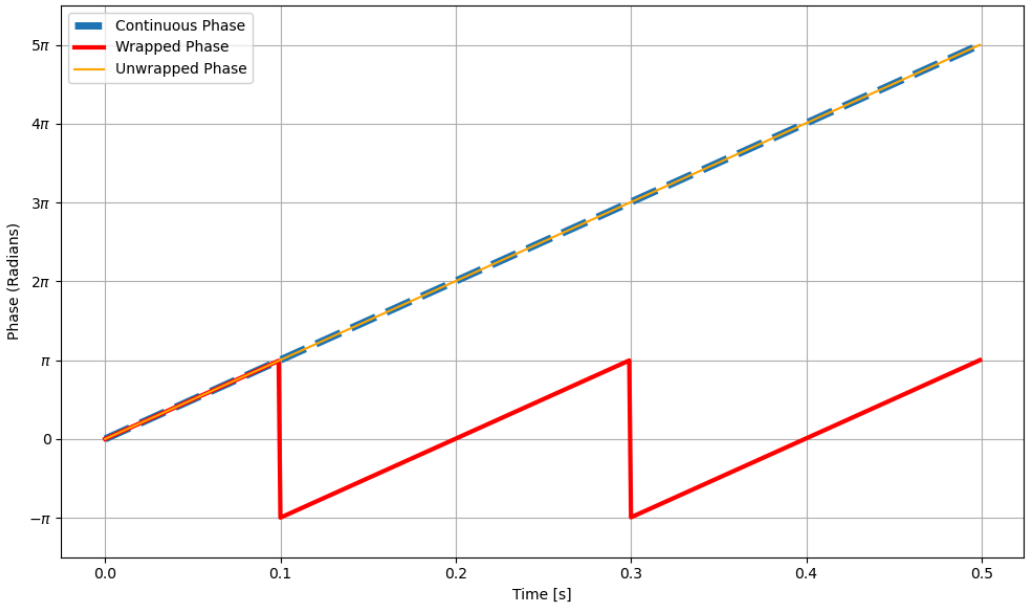

Figure 3.2: Illustration of the continuous, wrapped, and unwrapped phases.

3.1.1 Unwrapped instantaneous phase (UIP) feature

To determine the instantaneous phase from each sub-band signal of the analytic representation obtained from the output of an auditory filter bank, the deviation of the absolute phases from the center frequency of that sub-band signal can be calculated, as described in Eq. (3.5). However, due to the wrapping nature of the inverse tangent function during phase calculation, i.e., $\phi(r,t) \in [-\pi, \pi)$, it becomes challenging to interpret the characteristics of the phase trajectory along the time or frequency axes. Abrupt phase angle jumps at the wrapped points complicate this interpretation. Therefore, the unwrapping process is essential for reconstructing a continuous phase before performing any further derivative analyses. The illustrations of continuous phase, wrapped phase, and unwrapped phase are depicted in Figure 3.2.

This study leverages the phase unwrapping algorithms as follows [38]. Considering the time domain, the difference between adjacent elements in the phase trajectory is calculated as

$$\delta\phi(r,t) = \phi(r,t) - \phi(r,t-1), \tag{3.7}$$

where $\phi(r,t)$ in Radians is the wrapped phase value at time t.

Let ξ be the discontinuity threshold caused by phase wrapping. If $|\delta\phi(r,t)| > \xi$, the unwrapped phase difference is determined as follows

$$\delta\widetilde{\phi}(r,t) = \delta\phi(r,t) - 2\pi \,\text{round}\left(\frac{\delta\phi(r,t)}{2\pi}\right), \tag{3.8}$$

with round(x) is the round operation. The unwrapped phase value is then calculated as the cumulative sum of those unwrapped phase differences. The formula for this calculation is presented as follows

$$\widetilde{\phi}(r,t) = \widetilde{\phi}(r,t-1) + \delta\widetilde{\phi}(r,t), \tag{3.9}$$

where $\widetilde{\phi}(r,t_0) = \phi(r,t_0)$ and $\widetilde{\phi}(r,t)$ is the unwrapped instantaneous phase spectrogram along the temporal axis at start time t_0 and time t.

Additionally, this algorithm is also similarly applied along the frequency axis to derive the unwrapped phase spectrogram along the frequency axis, i.e, $\widetilde{\theta}(r,t)$, instead of the temporal axis.

3.1.2 Time derivative of phase feature

Time derivative of phase (TDP), also referred to as instantaneous frequency [41], 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:

$$f(r,t) = \frac{1}{2\pi}\frac{\partial\widetilde{\phi}(r,t)}{\partial t}, \tag{3.10}$$

with $\widetilde{\phi}(r,t)$ in Radians is the unwrapped instantaneous phase along 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 [42], the value of TDP can be es-

timated as follows

$$\frac{\partial\widetilde{\phi}(r,t)}{\partial t} = \widetilde{\phi}(r,t+\Delta t) - \widetilde{\phi}(r,t), \tag{3.11}$$

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

3.1.3 Frequency derivative of phase feature

Frequency derivative of phase (FDP), also known as group delay [41], 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:

$$\tau_g(r,t) = -\frac{\partial\widetilde{\theta}(r,t)}{\partial r}\frac{\partial r}{\partial\omega}, \tag{3.12}$$

with $\widetilde{\theta}(r,t)$ in Radians is the unwrapped phase spectrogram along 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 [42], FDP can be approximated numerically as

$$\frac{\partial\widetilde{\theta}(r,t)}{\partial r}\frac{\partial r}{\partial\omega} = \frac{\widetilde{\theta}(r+\Delta r,t) - \widetilde{\theta}(r,t)}{\omega_{r+\Delta r} - \omega_r}, \tag{3.13}$$

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

3.1.4 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, the temporal differentiation operation is first applied 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

$$\frac{\partial^2 \widetilde{\phi}(r,t)}{\partial \omega \partial t} = -\frac{\partial(\widetilde{2\pi f}(r,t))}{\partial r} \times \frac{\partial r}{\partial \omega}. \tag{3.14}$$

where $2\pi f(r,t)$ spectrogram is unwrapped along the frequency axis to obtain $\widetilde{2\pi f}(r,t)$ before performing frequency differentiation.

In the discrete-time domain, TFDP can be numerically approximated by employing the finite difference method, following the approach outlined in Eq. (3.11) and Eq. (3.13).

3.2 Validating the derivation using frequency modulation signal

This section outlines the procedure for validating the correctness of all derivation steps related to the proposed phase-based features. A frequency modulation signal is simulated and injected with various artificial interruptions to its instantaneous phase deviation. Following this, a GTFB is applied to the simulation for phase analysis. The instantaneous phase information of each output sub-band signal is observed, and the time/frequency derivative is calculated. If these derivatives successfully reflect the interruptions, this phase analysis method can be applied to detect the phase interruption in anomalous sound.

Frequency modulation and phase interruption

Frequency modulation (FM) is a technique used to encode information in a carrier signal by varying its instantaneous frequency in proportion to the amplitude of the message signal [41]. For a single-tone FM signal with a sinusoidal carrier, the formula is as follows:

$$x(t) = A_c \cos\left(2\pi f_c t + 2\pi f_\Delta \int_0^t Q(\tau) d\tau\right), \tag{3.15}$$

where A_c is the amplitude of carrier signal, f_c is the carrier frequency, f_Δ is the frequency deviation and $Q(t)$ is the message signal modulating the

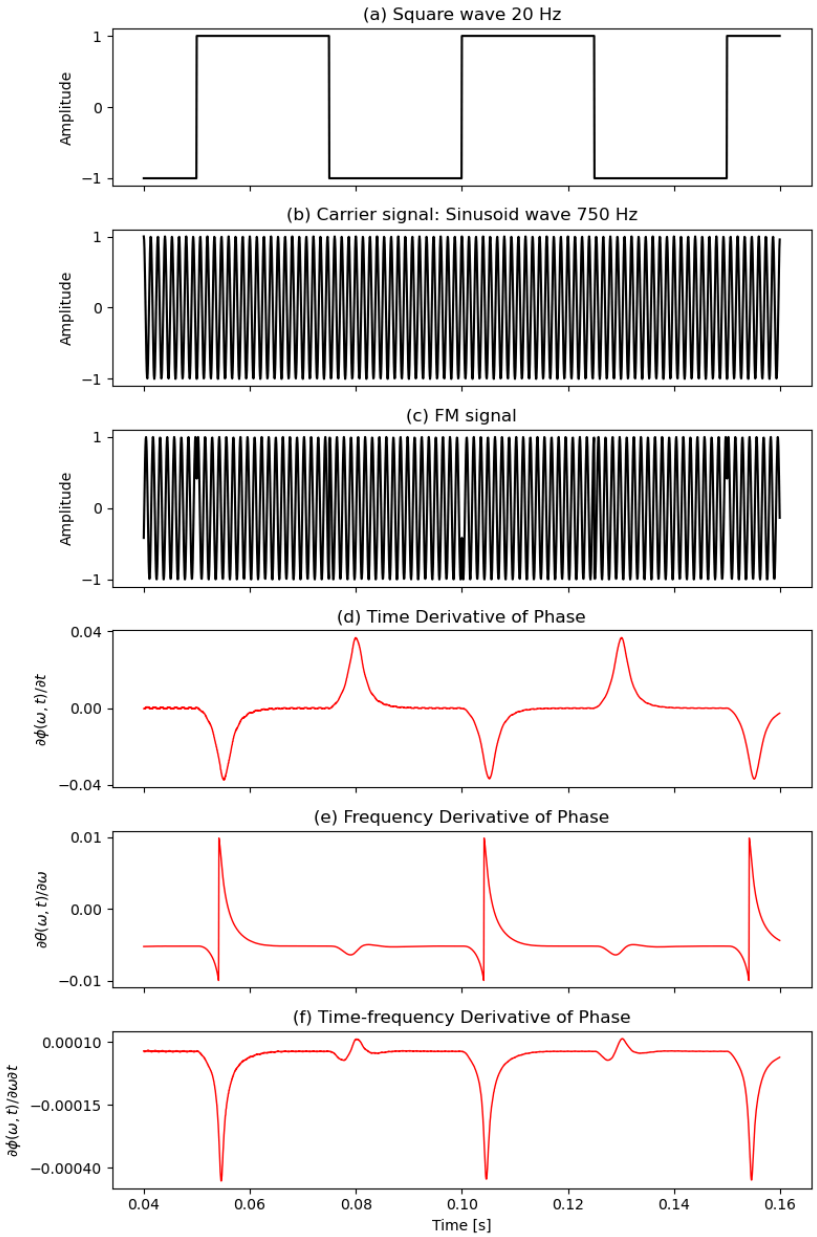

Figure 3.3: Simulation results of phase analysis for an FM signal utilizing a GTFB: (a) the instantaneous phase of FM signal, (b) the carrier signal in the time domain, (c) the FM signal, (d), (e), and (f) represent the time, frequency, and time-frequency derivatives of the phase of the sub-band signal centered at 750 Hz, respectively.

carrier. Additionally, the instantaneous phase of FM signal is defined as

$$\phi(t) = 2\pi f_\Delta \int_0^t Q(\tau)d\tau. \tag{3.16}$$

To simulate phase interruptions, an impulse train—also referred to as a train of Dirac delta functions—is used as the message signal in the frequency modulation process. The spikes in the impulse train signal can lead to abrupt changes in the instantaneous frequency during modulation, causing a phase interruption. In this scenario, the instantaneous phase resembles a square wave after calculating the anti-derivative of the impulse train signal.

Simulation procedure

Using the single-tone FM signal, the simulation procedure is as follows:

1. Input the following parameters: f_c, f_m, and f_Δ,
2. Construct a single-tone FM signal $x(t)$ with a message signal $Q(t)$, where $Q(t)$ is the impulse train signal. In this scenario, the instantaneous phase of $x(t)$ behaves like a square wave with frequency f_m,
3. Using a time-domain GTFB to filter $x(t)$,
4. Using Hilbert transform to represent the filtered output as an analytic representation,
5. Calculate TDP, FDP and TFDP as presented in (3.11), (3.13) and (3.14).

Simulation results

The configuration for the simulation in this section is as follows: The carrier signal is sinusoidal with a frequency of $f_c = 750$ Hz. A square wave with a frequency of 20 Hz varies the instantaneous phase of the FM signal. The instantaneous frequency deviation of the FM signal is $f_\Delta = 40$ Hz, and the overall sampling rate for the simulation is $f_s = 16000$Hz.

Figure 3.3 illustrates the simulation of the FM signal and the corresponding phase analysis using a GTFB. The plots displaying the TDP, FDP, and TFDP of the sub-band signal at a center frequency of 750 Hz are provided

for observation. This visualization shows that TDP, FDP, and TFDP can effectively highlight the artificial interruptions in the simulated instantaneous phase. The results of this simulation confirm the feasibility of the proposed phase analysis method. The next section will outline the effectiveness validation process for ASD by utilizing phase-based features.

3.3 Validating the effectiveness of phase-based features for ASD via supervised learning

This section outlines the procedure for validating the effectiveness of using phase-based features in ASD through supervised learning. The supervised learning algorithm for classifying normal instances and anomalies is implemented in experiments using the MIMII dataset. The overall performance of the supervised ASD task that employs phase-based features is assessed using accuracy, F1-score, and Matthews Correlation Coefficient. High values in both metrics indicate the effectiveness of the proposed feature in detecting anomalous sound.

3.3.1 Validation setup

Dataset preparation

The validation procedure conducts an experiment on the well-known MIMII dataset [3] with SNR = 6 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 anomalous 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 a supervised manner using both normal and anomalous data. Inference processes are also conducted with both normal and anomalous data to evaluate overall performance. The statistics of normal and anomalous sound in MIMII are described in Table 3.1.

Figure 3.4 (a) illustrates the dataset preparation process. Each section corresponding to a specific machine type ID, such as Slider 00, is shuffled and

Table 3.1: 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

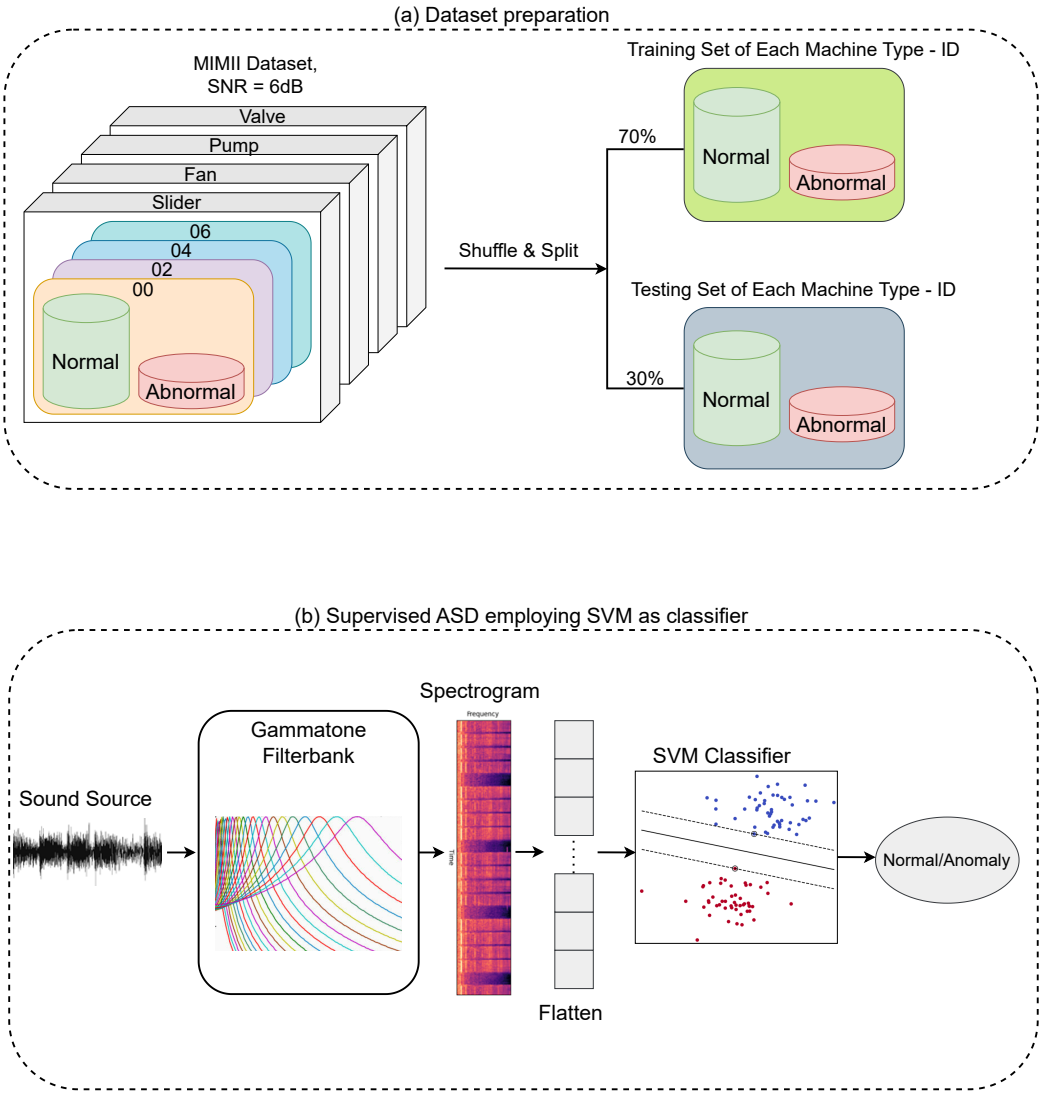

Figure 3.4: Illustration of validation procedure, including dataset preparation and supervised ASD employing SVM as classifier.

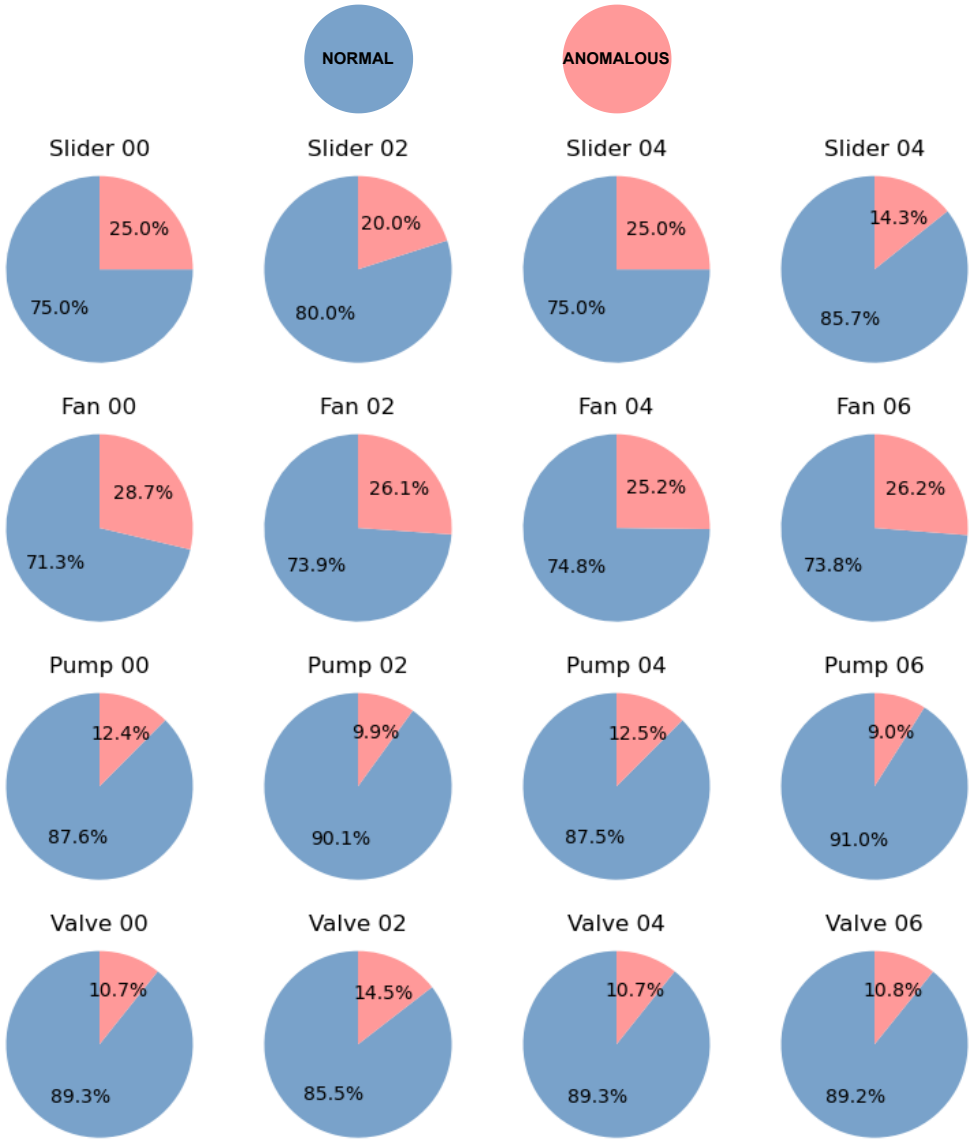

Figure 3.5: The percentage ratio between normal data and abnormal data in the MIMII dataset at SNR = 6 dB.

divided into training and testing sets with a 70:30 ratio. The ratio between normal and anomalous data is maintained consistently in both the training and testing sets, mirroring the ratio found in the original dataset. Figure 3.5 illustrates the percentage ratio between normal data and anomalous data in the MIMII dataset at SNR = 6 dB.

Anomalous sound classification

This research leverages Support Vector Machine (SVM) [44] as the binary classification system. With the robustness and effectiveness of a supervised learning algorithm, SVM is particularly well-suited for classification tasks involving high-dimensional data. To utilize the SVM for binary classification, normal sounds are labeled as negative (-1) and anomalous sounds as positive (1) to establish the ground truth for training and testing. The features of UIP, 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 flattened and subsequently fed into the SVM model. This procedure is depicted in Figure 3.4 (b).

Evaluation metrics

This study uses accuracy, F1-score, and MCC to evaluate supervised ASD performance in terms of the classification task by SVM. Accuracy measures the percentage of correctly predicted instances compared to the total number of predictions. Accuracy is calculated as follows

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{3.17}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

F1-score is a metric specifically designed for balanced assessment of the model's performance. It is calculated as the harmonic mean of precision and

recall:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3.18}$$

where precision represents the proportion of true positive class with total predicted positive samples. Meanwhile, recall stands for sensitivity. The formula of precision and recall is expressed as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{3.19}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3.20}$$

In the case of an imbalanced dataset, such as in anomaly detection tasks, accuracy is often skewed towards the majority class. Therefore, Matthew's Correlation Coefficient (MCC) is frequently utilized to evaluate system performance. The calculation formula is

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \tag{3.21}$$

3.3.2 Results

Tables 3.2, 3.3, and 3.4 present the performance results evaluated in terms of accuracy, F1-score, and MCC metrics for supervised ASD using SVM classification and phase-based features. The outperformed results are in bold text, and the results of target machines in the research issue, including Slider (ID 06) and Valve (ID 06) are colored. The experimental results indicate the following:

1. The classification performance using phase-based features highlights the superiority of supervised ASD systems in detecting anomalous sounds from sliders, fans, and pumps. The accuracy and F1-score results for these ASD systems are predominantly above 0.9, demonstrating the strong ability of phase-based features in distinguishing between normal and anomalous sounds emitted by these machines.

2. Furthermore, the MCC results shown in Table 3 illustrate that the supervised ASD systems for sliders, fans, and pumps, which utilize

phase-based features and SVM classification, perform well even with imbalanced datasets. Most of the MCC results are close to 1.0, achieved without the use of over-sampling methods for data augmentation.

3. Additionally, the analysis of the supplementary role of the derivative of phase across time and/or frequency axes reveals that combining both axes can further enhance the overall performance in detecting anomalous sounds, such as those from fans, with all metrics—accuracy, F1-score, and MCC—reaching 1.0.

4. The performance results highlight the challenges faced by SVM in detecting abnormal valve sounds when using phase-based features. This is evident in the lower F1-score and MCC compared to other machine learning models, despite achieving high overall accuracy. Focusing specifically on the valve machine type, the performance of the ASD task for valves 02 and 04 is superior to that of the others, with the MCC exceeding 0.6.

To better understand the model's predictions for valve performance, confusion matrices for ASD are provided in Figure 3.6, 3.7, 3.8, and 3.9. These matrices indicate that:

1. The ASD model using SVM, with UIP as the input feature, shows improved performance in detecting anomalous sounds from the valve compared to other models using phase. This is evidenced by a higher true positive ratio, which treats anomalies as the positive class. However, the model does have a higher false positive ratio for the negative class, which represents normal sounds.

2. Furthermore, the ASD model using SVM and TDP as input features shows a slight improvement in performance for valve 02, as the false positive ratio is reduced while maintaining a balance in the true positive ratio compared to SVM+UIP.

3. Additionally, the ASD model using SVM and FDP/TFDP as input features performs the worst in detecting anomalous sounds from valves, even though the ratio of false positives has been significantly reduced.

Table 3.2: Accuracy results of supervised ASD utilizing SVM and phase-based features.

Machine Type	ID	Accuracy			
		UIP	TDP	FDP	TFDP
Slider	00	0.998	**1.000**	**1.000**	**1.000**
	02	0.995	0.995	0.995	0.993
	04	**1.000**	**1.000**	0.995	0.986
	06	0.995	0.995	0.984	0.984
Fan	00	0.995	0.995	0.998	**1.000**
	02	0.998	0.998	**1.000**	**1.000**
	04	0.997	**1.000**	**1.000**	**1.000**
	06	**1.000**	**1.000**	0.993	**1.000**
Pump	00	0.991	0.986	0.980	0.994
	02	**1.000**	**1.000**	0.994	0.994
	04	**1.000**	**1.000**	0.996	0.996
	06	0.994	0.994	0.991	0.994
Valve	00	0.916	0.931	0.925	0.919
	02	0.948	0.948	0.924	0.924
	04	0.935	0.940	0.946	0.935
	06	0.895	0.925	0.904	0.901
Average on each machine type					
Slider		0.997	0.998	0.994	0.991
Fan		0.998	0.998	0.998	**1.000**
Pump		0.996	0.995	0.991	0.995
Valve		0.924	0.936	0.925	0.920

Table 3.3: F1-score results of supervised ASD utilizing SVM and phase-based features.

Machine Type	ID	F1-score			
		UIP	TDP	FDP	TFDP
Slider	00	0.995	**1.000**	**1.000**	**1.000**
	02	0.987	0.987	0.987	0.981
	04	**1.000**	**1.000**	0.990	0.971
	06	0.981	0.981	0.941	0.941
Fan	00	0.992	0.992	0.994	**1.000**
	02	0.995	0.995	**1.000**	**1.000**
	04	0.995	**1.000**	**1.000**	**1.000**
	06	**1.000**	**1.000**	0.986	**1.000**
Pump	00	0.964	0.938	0.911	0.976
	02	**1.000**	**1.000**	0.969	0.969
	04	**1.000**	**1.000**	0.983	0.983
	06	0.967	0.967	0.949	0.967
Valve	00	0.674	0.610	0.490	0.400
	02	0.817	0.794	0.655	0.655
	04	0.686	0.655	0.679	0.560
	06	0.615	0.638	0.273	0.154
Average on each machine type					
Slider		0.991	0.992	0.980	0.973
Fan		0.996	0.997	0.995	**1.000**
Pump		0.983	0.976	0.953	0.974
Valve		0.698	0.674	0.524	0.442

Table 3.4: MCC results of supervised ASD utilizing SVM and phase-based features.

Machine Type	ID	MCC			
		UIP	TDP	FDP	TFDP
Slider	00	0.994	**1.000**	**1.000**	**1.000**
	02	0.984	0.984	0.984	0.976
	04	**1.000**	**1.000**	0.987	0.962
	06	0.978	0.978	0.934	0.934
Fan	00	0.989	0.989	0.994	**1.000**
	02	0.994	0.994	**1.000**	**1.000**
	04	0.994	**1.000**	**1.000**	**1.000**
	06	**1.000**	**1.000**	0.982	**1.000**
Pump	00	0.960	0.923	0.905	0.973
	02	**1.000**	**1.000**	0.966	0.966
	04	**1.000**	**1.000**	0.981	0.981
	06	0.964	0.964	0.946	0.964
Valve	00	0.639	0.592	0.529	0.479
	02	0.787	0.775	0.656	0.656
	04	0.650	0.647	0.686	0.602
	06	0.575	0.597	0.324	0.274
Average on each machine type					
Slider		0.989	0.991	0.976	0.968
Fan		0.994	0.996	0.994	**1.000**
Pump		0.981	0.972	0.950	0.971
Valve		0.663	0.653	0.549	0.503

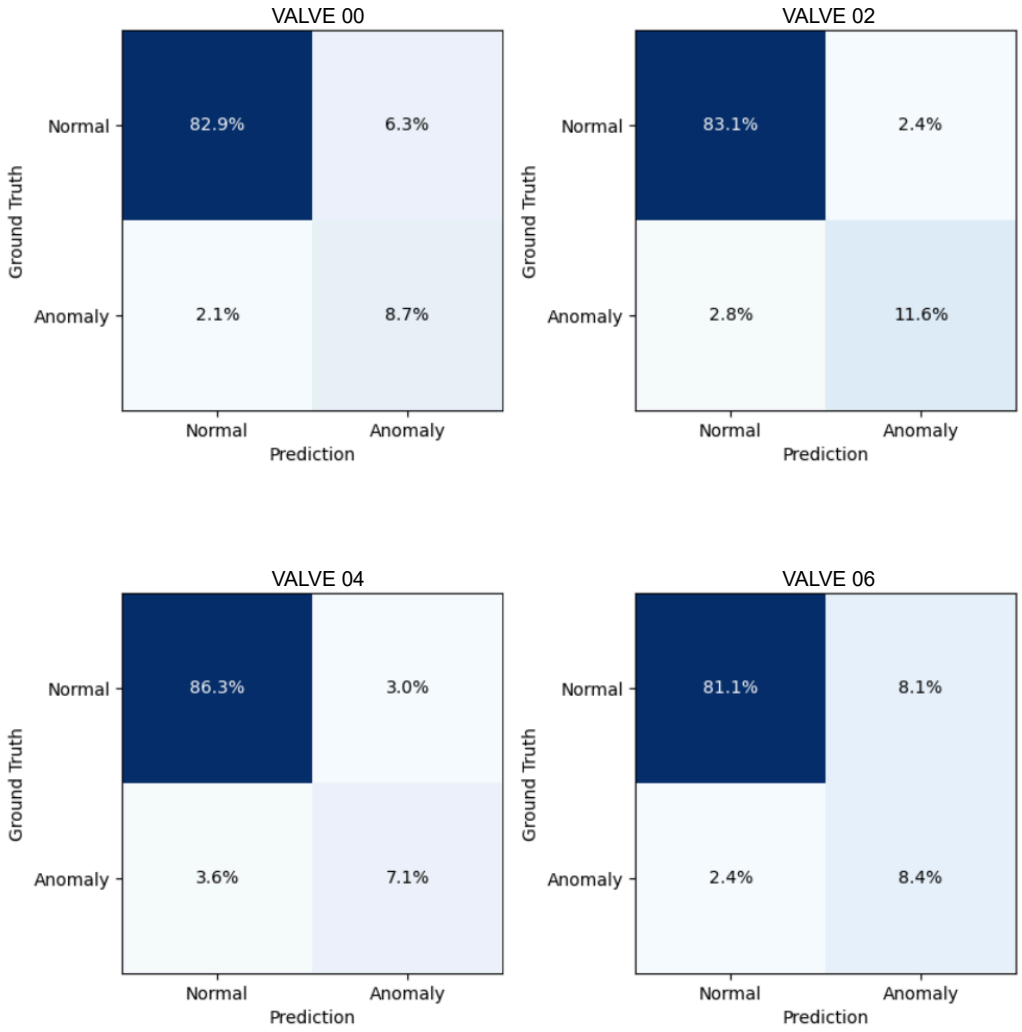

Figure 3.6: Confusion matrix results for supervised ASD from Valve using SVM+UIP.

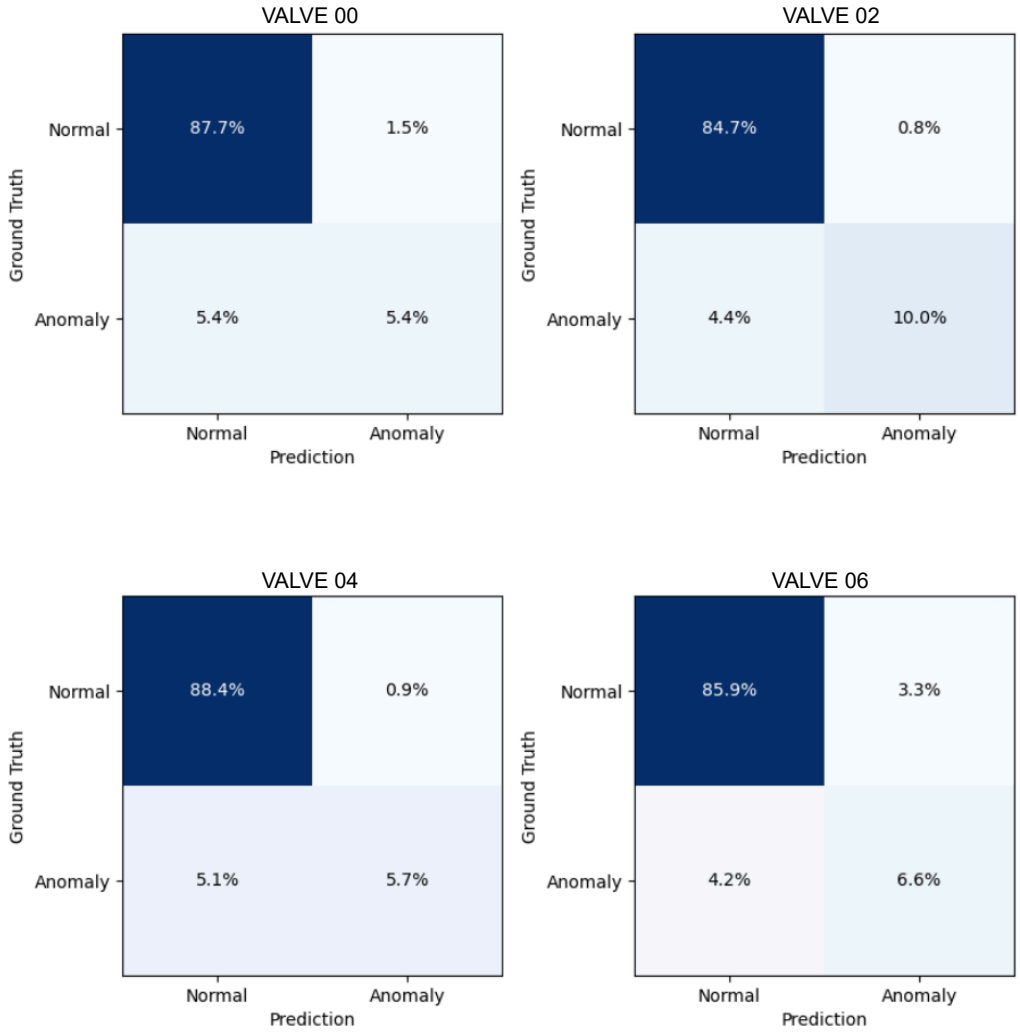

Figure 3.7: Confusion matrix results for supervised ASD from Valve using SVM+TDP.

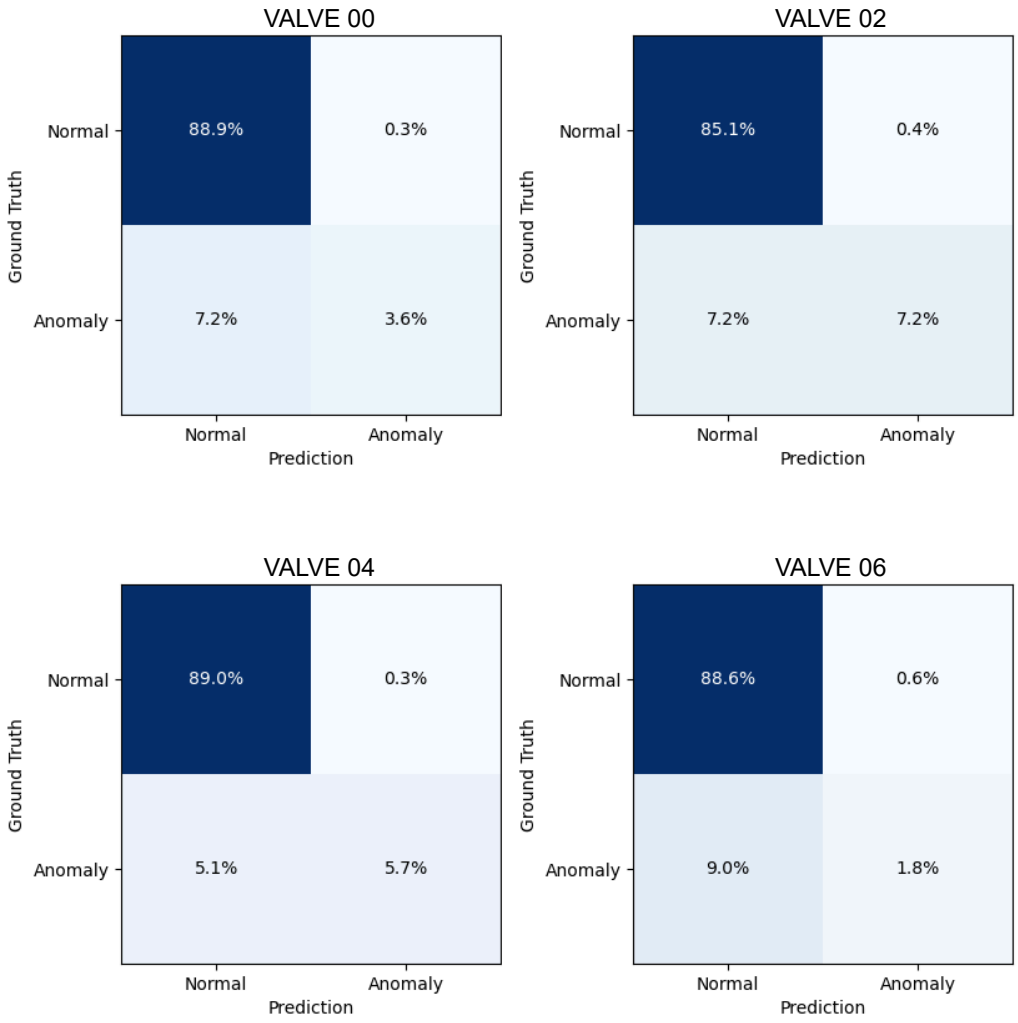

Figure 3.8: Confusion matrix results for supervised ASD from Valve using SVM+FDP.

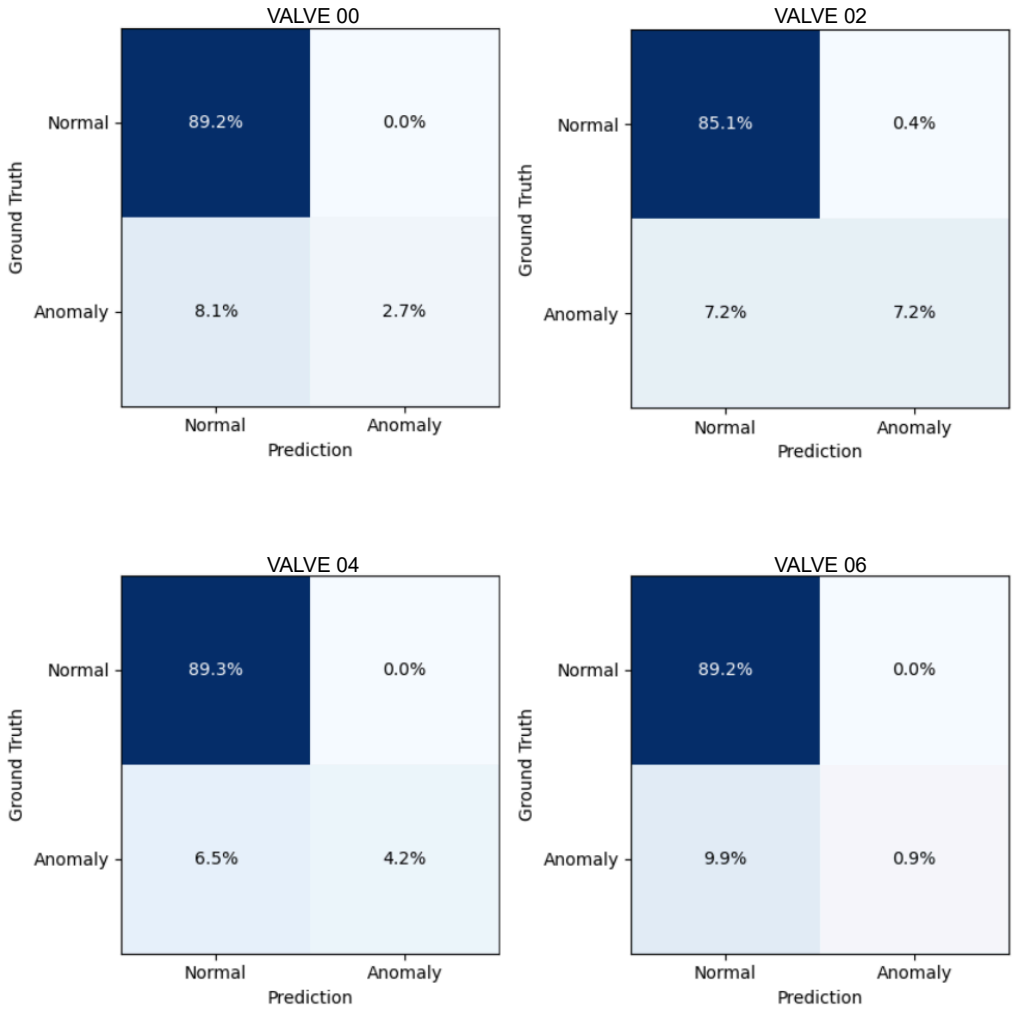

Figure 3.9: Confusion matrix results for supervised ASD from Valve using SVM+TFDP.

3.3.3 Discussion

The evaluation results from the supervised ASD using SVM classification with phase-based features provide strong evidence that these features can effectively detect anomalous sounds from industrial machines. The validation demonstrates that phase-based features can capture distinct patterns in phase interruptions associated with anomalous sounds, particularly during periods of increased friction. This includes recognizing anomalous sounds from Slider and Fan, as well as detecting clogs in Pump sounds. Additionally, the experimental results indicate that phase-based features can identify abnormal Valve sounds in certain cases; however, their performance is generally less comparable to that of other machines. Moreover, the oversampling method [43] is also applied to Valve data before applying SVM; however, the performance could not be improved.

Chapter 4

Unsupervised ASD utilizing phase-based features

Anomalous sounds resulting from mechanical faults are referred to as zero-resource data due to the difficulties in acquisition and labeling. This situation highlights the necessity for an unsupervised anomaly detection system that does not require training with anomalous data. This chapter presents an unsupervised model that utilizes phase-based features for ASD to address the second research issue. The system's performance is evaluated through experiments conducted on the MIMII dataset.

4.1 Unsupervised ASD model using phase-based features

An autoencoder is commonly used as a baseline for unsupervised anomaly detection across various fields [30]. An autoencoder consists of two key components: the encoder and the decoder. The model is designed to learn complex patterns in data by encoding these patterns into a latent space. The decoder then reconstructs the latent features. This interaction between the encoder and decoder enhances the robustness of both components. The autoencoder is trained exclusively on normal data, allowing it to capture the characteristics of this data without exposure to anomalies. Anomalies can then be identified by examining the reconstruction error, which is expected

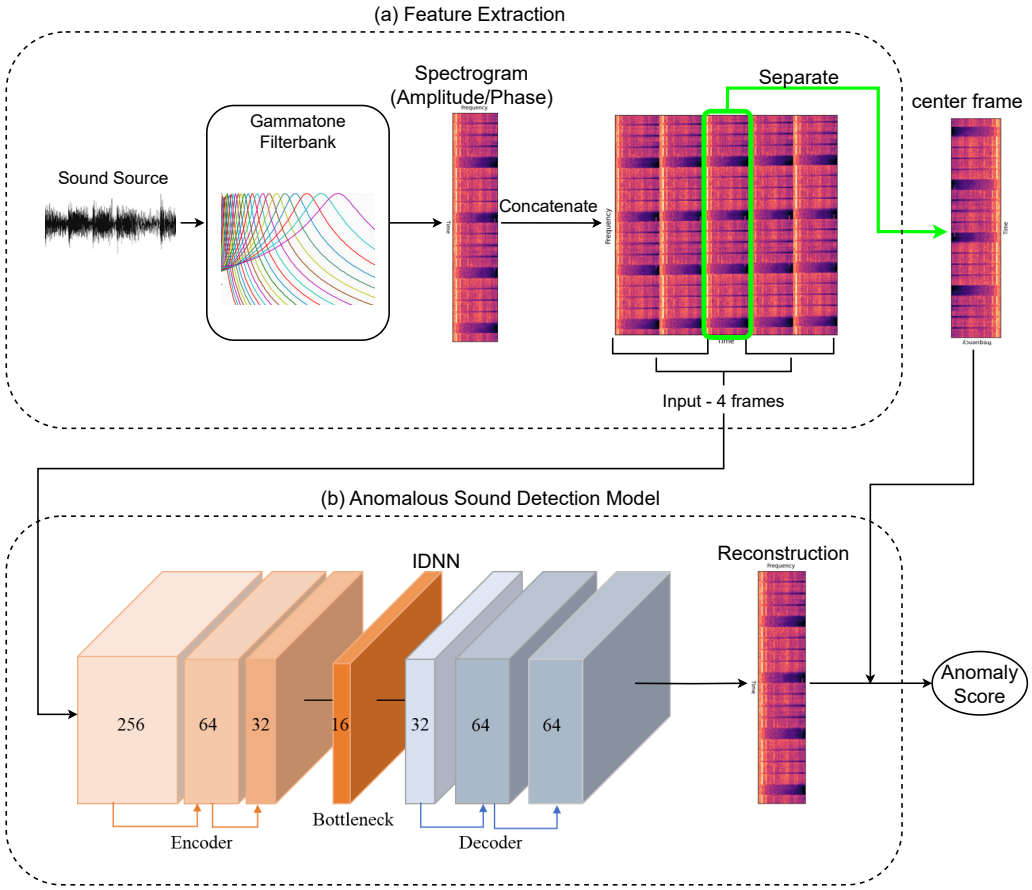

Figure 4.1: Illustration of ASD based on spectrograms employing an IDNN model. The input sound is filtered with a GTFB to extract amplitude or phase-based features, which are then represented as spectrograms.

to be significant for anomalous data. In this work, an Autoencoder is utilized as the backbone for the anomaly detection system, using phase-based features as inputs.

4.1.1 Phase-based feature extraction

Four phase-based features serve as inputs for the unsupervised ASD model. These features include UIP, TDP, FDP, and TFDP. They are extracted from the output of a time-domain GTFB with center frequencies distributed linearly on the ERB scale, which are presented in Chapter 3. The features are represented as two-dimensional spectrograms, capturing both time and frequency information, which facilitates the spectrogram-based learning process of the Autoencoder-based ASD model. Additionally, the spectrograms obtained after filtering the input sounds are then preprocessed to better match the input size required by the ASD model. For example, the input for the Autoencoder-based model in the ASD baseline [5] is usually constructed by concatenating five consecutive frames over time, which are then fed into the model. Figure 4.1(a) illustrates the phase-based feature extraction process.

4.1.2 Autoencoder-based 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-based Interpolation Deep Neural Network (IDNN) has been proposed [39]. 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

$$\mathcal{L}\left(x_{\frac{n+1}{2}} \middle| \mathcal{D}(\mathcal{E}([x_1, \ldots, x_{\frac{n+1}{2}-1}, x_{\frac{n+1}{2}+1}, \ldots, x_n]))\right), \tag{4.1}$$

where $\mathcal{E}, \mathcal{D}$ and $\mathcal{L}$ are the encoder, decoder, and loss function used in IDNN.

In this study, IDNN is utilized 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 Figure 4.1.

This study utilizes the Mean Square Error (MSE) function as the loss function for training the IDNN. To evaluate the reconstruction error, the reconstructed spectrogram output from the IDNN is compared to the middle frame using MSE. Additionally, this error value is employed as the anomaly score calculated from the ASD model. The formula for calculating MSE is presented as follows

$$\mathrm{MSE} = \frac{1}{\mathrm{T} \times \mathrm{F}} \sum_{t=1}^{\mathrm{T}} \sum_{f=1}^{\mathrm{F}} (S_{t,f} - \hat{S}_{t,f})^2, \tag{4.2}$$

where $S, \hat{S} \in \mathbb{R}^{\mathrm{T} \times \mathrm{F}}$ are the original spectrogram and reconstructed spectrogram obtained from IDNN. Additionally, T, F represents the shape of spectrogram as Time×Frequency.

4.2 Implementation

The features of UIP, TDP, FDP, and TFDP are extracted using a time-domain FIR GTFB. The center frequencies in the filterbank are distributed linearly along the ERB scale, ranging from 2 Cam to 32 Cam. To reduce the temporal dimension, the resulting spectrograms are downsampled 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 across five frames to form 320-dimensional input vectors. The middle frame is designated as the target spectrogram, while the frames on the left and right are concatenated to serve as the input for the model. All models are trained concurrently for 200 epochs, with a batch size of 64, utilizing the Adam optimizer [46] with a learning rate of 0.001, and mean squared error (MSE) to calculate the reconstruction error. Details of the employed IDNN model are presented in Table 4.1.

Table 4.1: 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

Chapter 5

Evaluation

5.1 Dataset preparation

The assessment of unsupervised ASD using IDNN-based models and phase-based features is performed on the MIMII dataset [3]. Similar to the evaluation of the previous supervised system, this process also involves shuffling and dividing the dataset into training and testing sets, although there are some notable differences. Each data section labelled by machine type - machine ID, e.g., slider 00, in the dataset contains both anomalous and normal data, with the percentage ratio illustrated in Figure 2. For each data section, the training set includes only normal data from that machine, while the testing set consists of both normal and anomalous data, featuring an equal number of samples from each category. The unsupervised ASD models are trained exclusively on the normal data in the training set, and their overall performance is assessed on the testing set. Furthermore, no data augmentation techniques will be applied in these experiments.

5.2 Evaluation metrics

Anomaly detection functions as a classification task with two distinct classes. Utilizing an unsupervised learning approach, the final determination is evaluated based on anomaly score thresholding. However, selecting an appropriate threshold for ASD can be pretty challenging, leading most thresholds to be

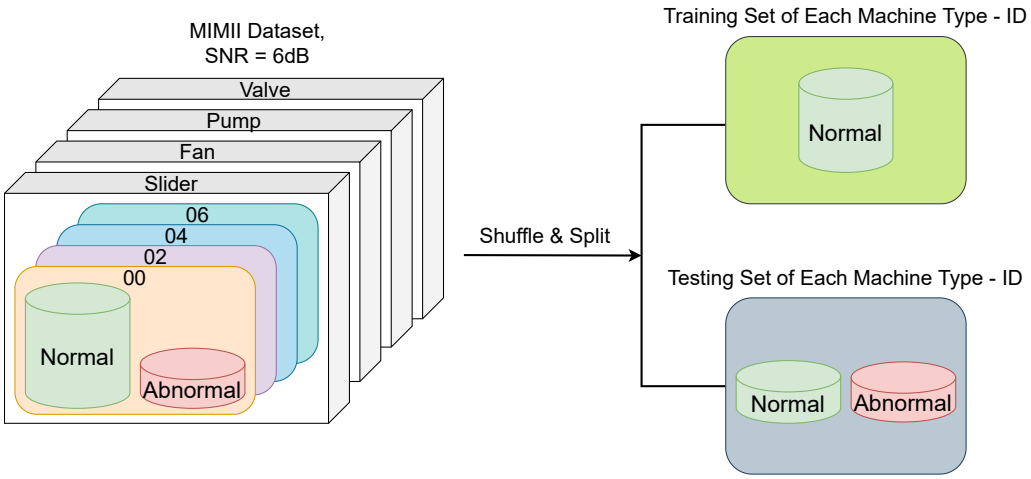

Figure 5.1: Dataset preparation for training and testing the unsupervised ASD system.

defined heuristically [35]. This study utilizes the Area Under the Receiver Operating Characteristic curve (AUC-ROC) [45] 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

$$\mathrm{AUC} = \frac{1}{\mathrm{N}_{-}\mathrm{N}_{+}} \sum_{\mathrm{i}=1}^{\mathrm{N}_{-}} \sum_{\mathrm{j}=1}^{\mathrm{N}_{+}} \mathcal{H}(\mathcal{A}_{\theta}(\mathrm{x}_{\mathrm{j}}^{+}) - \mathcal{A}_{\theta}(\mathrm{x}_{\mathrm{i}}^{-})), \tag{5.1}$$

where $\{\mathrm{x}_{\mathrm{i}}^{-}\}_{\mathrm{i}=1}^{\mathrm{N}_{-}}$ and $\{\mathrm{x}_{\mathrm{j}}^{+}\}_{\mathrm{j}=1}^{\mathrm{N}_{+}}$ are the normal and anomalous test samples, with N_{-} normal samples and 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.

5.3 Comparative performance evaluation

Table 5.1 presents the performance results of the IDNN model based on the AUC score, utilizing five different features. The names of the proposed features, including UIP, TDP, FDP, and TFDP, are highlighted in the table. Improvements compared to instantaneous amplitude (IA) are indicated in

bold text.

The IDNN models that utilize UIP, TDP, and TFDP features demonstrate superior effectiveness in detecting anomalous sounds from machinery, specifically sliders, fans, and pumps, compared to the IA feature used in ASD. Notably, for the pumps, there is an approximate 5% and 8.1% increase in the AUC score when employing TDP/TFDP and UIP as discriminative features compared to IA. Additionally, the models leveraging TDP and TFDP features outperform both IA and UIP in detecting abnormal sounds from the fan. The empirical results for the fans 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 sliders, abnormal sounds can also be effectively detected using either TDP or TFDP features, with TDP showing a slight improvement over both IA and UIP. The empirical 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 IA features, which remain the most effective for identifying abnormal sounds from valves.

This study compares the performance of the proposed method, which utilizes IDNN and phase-based features, with that of recently developed unsupervised methods in Table 5.2. These methods include DCASE 2020 Baseline [5], CVAE [48], GRLNet [49], IDNN+IA [24], and Deep SVDD [50]. These methods have used the amplitude information as the input feature. These results indicate that the proposed method outperforms others in AUC for detecting abnormal sounds from sliders, fans, and pumps. Additionally, it can be observed that the proposed method remains less competitive than the other methods in terms of performance in detecting anomalous sounds from valves.

To illustrate the effectiveness of the proposed phase-based features, this study performs t-SNE projection [47] to visualize the latent embedding of one of the machines in Figure 5.2. Additionally, the original features are visualized to eliminate the model's effect and better understand the behavior of the proposed features, as shown in Figure 5.3. By observing the black contours,

it can be confirmed 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 separate distinguishable clusters. Additionally, these visualizations acknowledge that the red contours highlight the slight challenges faced by the IDNN model in structuring phase-based information to its latent space in ASD, indicated by the concrete overlapping between normal and anomalous clusters.

5.4 Discussion

The evaluation results presented in this chapter highlight the effectiveness of phase information in detecting anomalous sounds from industrial machines, even without the necessity of training on anomalous data. These findings reinforce the study's hypothesis regarding the presence of interruption artifacts within the phase information of such sounds. Notably, the increase in the AUC score of ASD models, when utilizing the time-frequency derivative of phase compared to time or frequency derivatives, underscores the significance of considering both axes for a comprehensive detection approach. It is evident that phase-based features sometimes outperform amplitude-based features in ASD applications. This observation also points to the limitations of amplitude-based features in identifying anomalous sounds, as higher amplitudes may obscure sudden frequency changes, which are crucial for distinguishing anomalous sounds.

Besides, the experimental results also report the poor performance of unsupervised ASD using IDNN and phase-based features in detecting anomalous sound from valves. This poor performance may be attributed to the nature of the valve's sound, which is non-stationary and sparse over time. Additionally, the presence of silent segments in the valve sound can negatively impact the calculation of the reconstruction error between the target and the reconstructed spectrogram, ultimately compromising the effectiveness of the ASD models in detecting abnormal valve sounds.

Table 5.1: Performance comparison in AUC of unsupervised ASD employing five Gammatone features and IDNN-based models across different machines, with IA feature and the proposed features UIP, TDP, FDP, and TFDP.

Machine		Features				
		IA [24]	**UIP**	**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

Table 5.2: Comparison results in AUC of the proposed method and other unsupervised methods.

Authors	Methods		Machine Type			
	Model	Features	Slider	Fan	Pump	Valve
DCASE 2020 [5]	AE	Mel-spec	0.902	0.953	0.868	0.594
Nguyen et al. [48]	CVAE	Mel-spec	0.894	0.903	0.887	0.655
Yu et al. [49]	GRLNet		0.911	0.953	0.901	0.639
Hafiz et al. [24]	IDNN	IA	0.951	0.944	0.866	**0.931**
Kilickaya et al. [50]	Deep SVDD	Mel	0.867	0.936	0.837	0.804
Proposed	IDNN	UIP	0.950	0.948	**0.947**	0.660
	IDNN	TDP	**0.959**	0.956	0.913	0.752
	IDNN	FDP	0.668	0.904	0.807	0.696
	IDNN	TFDP	0.921	**0.981**	0.900	0.638

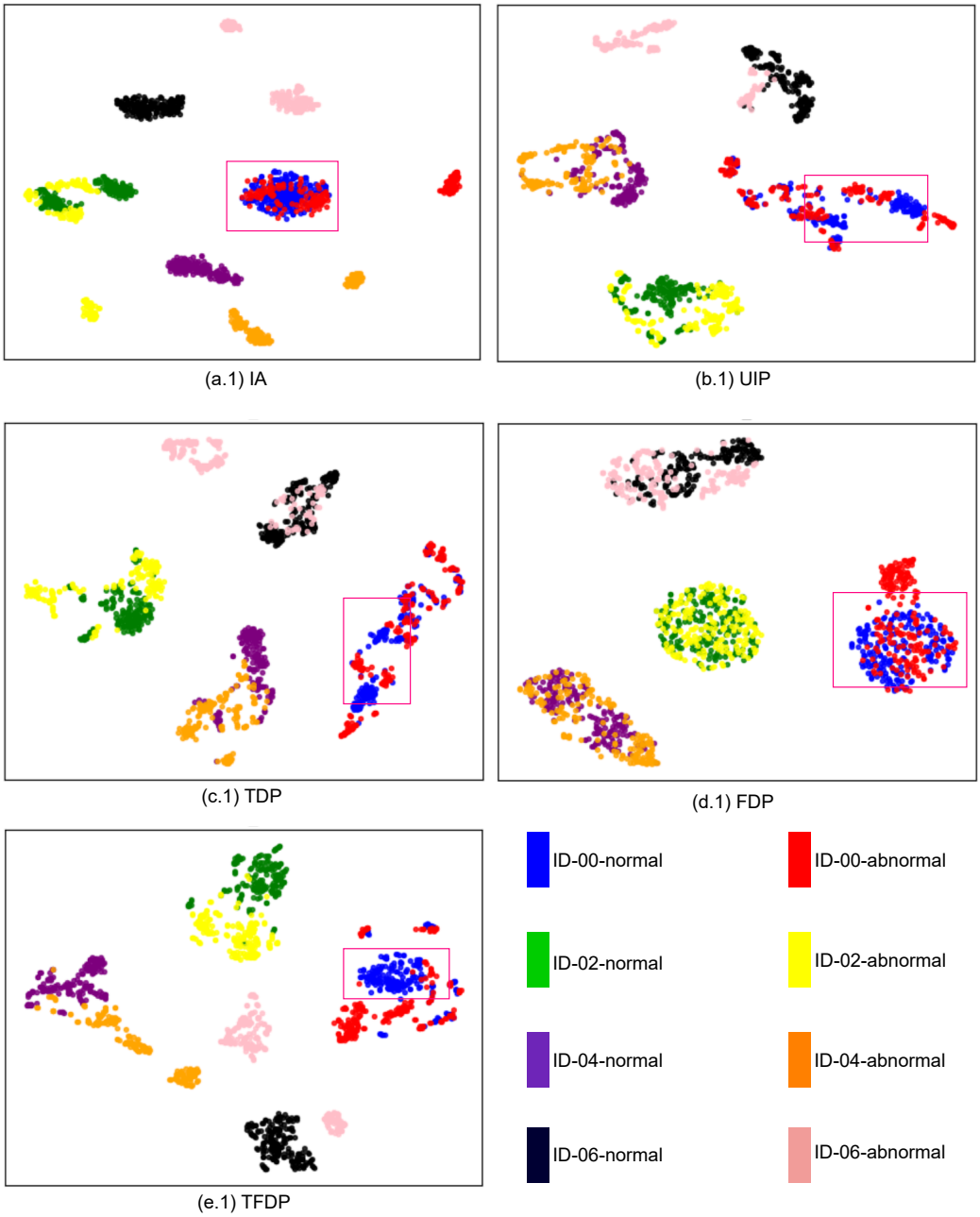

Figure 5.2: The t-SNE visualization of IDNN bottleneck features (a.1)–(e.1) of IA, UIP, TDP, FDP, and TFDP features of the Fan machine type in the MIMII dataset. Different colors represent different machine IDs and sound types. The pink contours demonstrate the significant discrimination ability of the proposed phase-based features in comparison with amplitude-based features.

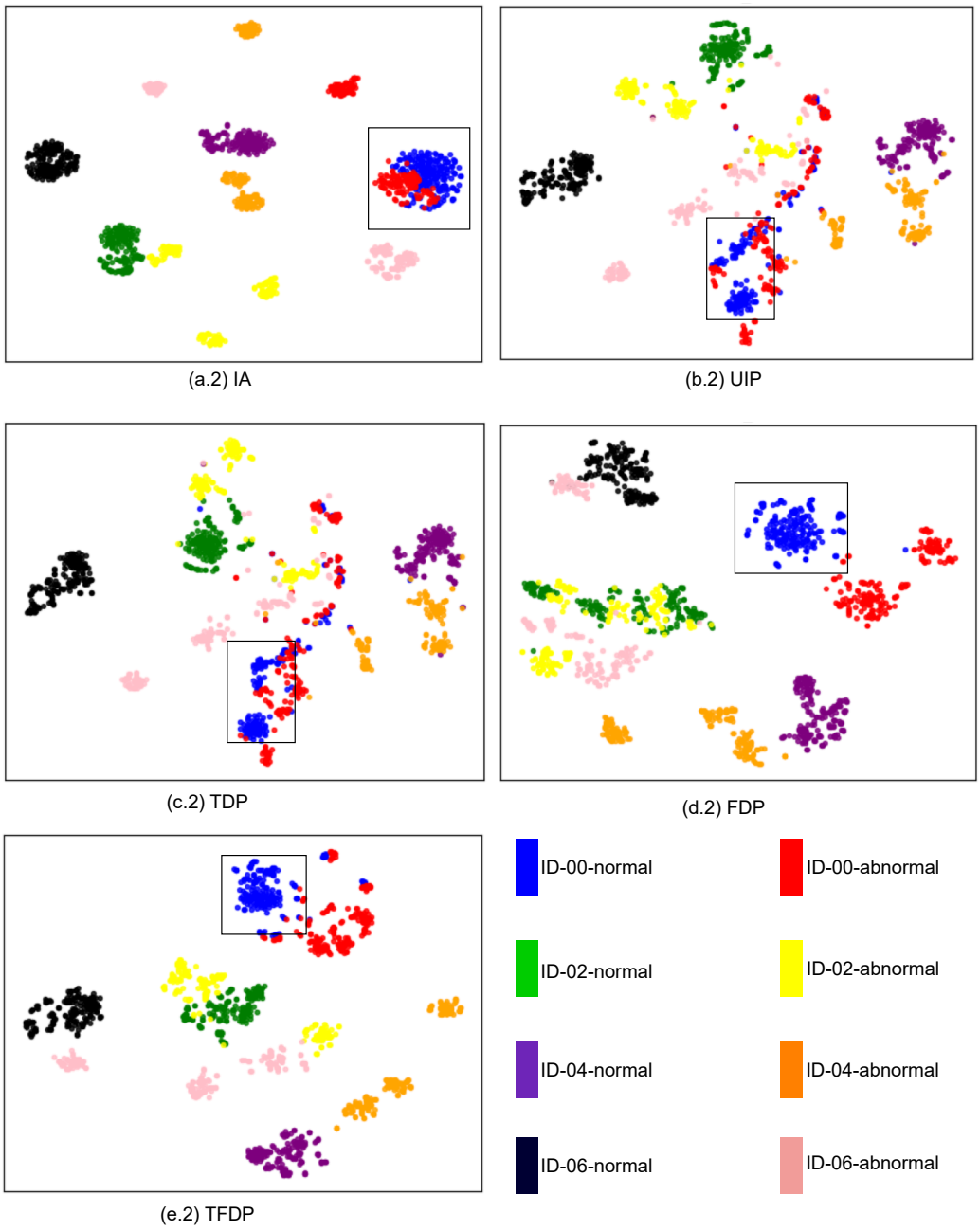

Figure 5.3: The t-SNE visualization of original features (a.2)–(e.2) of IA, UIP, TDP, FDP, and TFDP features of the Fan machine type in the MIMII dataset. Different colors represent different machine IDs and sound types. The black contours demonstrate the significant discrimination ability of the proposed phase-based features in comparison with amplitude-based features.

Chapter 6

Conclusion

6.1 Summary

This study focused on addressing the problem of anomalous sound detection (ASD), specifically the sounds emitted by industrial machines during malfunctions. The primary aim of the research was to propose a novel approach that utilizes instantaneous phase features for ASD, seeking to fill the research gap left by previous methods that relied on timbral attributes [2]. The motivation for this study arises from the hypothesis that phase interruptions exist in the acoustic characteristics of anomalous machine sounds. To achieve this goal, this study resolved the research question step by step, starting with the development of instantaneous phase features, then implementing and evaluating with a benchmark dataset to confirm the effectiveness of the proposed method.

First and foremost, this study developed the concepts of instantaneous phase features derived from the output of an auditory filterbank. To investigate the phase interruption of the hypothesis, the derivative of phase along time, frequency, and both axes was proposed. The developed features include unwrapped instantaneous phase (UIP), time derivative of phase (TDP), frequency derivative of phase (FDP), and time-frequency derivative of phase (TFDP). Secondly, to verify the derivation steps, a simulation with a frequency modulation signal containing artificial phase interruptions was also conducted. The simulation results showed that the derivative of phase

features could detect these interruptions, confirming the correctness of the proposed phase analysis method. Besides, a supervised ASD approach using Support Vector Machine (SVM) was conducted on the MIMII dataset [3] at SNR = 6 dB to validate the ability of phase-based features in detecting anomalous sound, evaluated with accuracy, F1-score, and Matthew Correlation Coefficient (MCC) metrics. The experimental results demonstrated that the proposed phase-based features work well in detecting anomalous sound from rotating machinery (Fan), sliding machinery (Slider), and liquid manipulator (Pump) in the MIMII dataset. Moreover, fusing both time and frequency derivatives of phase could enhance the overall performance in all machines. However, the poor performance in detecting anomalous valve sound using SVM and phase-based feature could be observed through this experiment.

This study also addressed the unsupervised ASD system utilizing instantaneous phase features to achieve the second research goal. By representing phase-based features as spectrograms with time and frequency axes, these spectrograms facilitated the integration with an unsupervised deep-learning-based model. This study leveraged Autoencoder-based Interpolation Deep Neural Network (IDNN) [39] as the backbone for an unsupervised ASD system, with the frontend being a phase-based features extractor. The experiment was conducted on the MIMII dataset at SNR = 6 dB and evaluated with the Area Under the Receiver Operating Characteristic curve (AUC). The experimental results showed that the unsupervised ASD using IDNN and phase-based features remains effective in detecting anomalous sound from Slider, Fan, and Pump, and outperforms other unsupervised methods that utilize amplitude information with these machine types.

6.2 Contributions

The main contribution of this work lies in both the application and scientific aspects. In the application aspect, this study helps develop an ASD detection system across industries, enhancing safety, security, and quality by identifying unusual sounds that may signal irregular events or malfunctions. From a scientific perspective, this study contributes to the research in au-

dio signal processing by utilizing instantaneous phase information, including proposing and validating the derivation for four representations of phase in both conceptual and implementation perspectives, whose full potential has yet to be explored in previous research. Additionally, these features can serve as a preliminary technique for other research, such as deepfake speech detection, phase-aware speech enhancement, etc., thereby improving accuracy, interpretability, and performance.

6.3 Future works

Despite the positive results reported from the experiments, several issues were identified. Firstly, the ASD system that uses phase-based features still struggles to detect anomalous sounds from valves. Valve sounds are non-stationary and occur sporadically over time, with segments of the audio clip often silent. This characteristic can confuse the ASD model, especially if the normal and abnormal sections in the dataset share this property.

Secondly, the AE-based unsupervised model tends to focus on minimizing reconstruction loss using normal data, which can lead to overfitting and degraded performance. Phase-based features carry rich information in both time and frequency domains, and may typically belong to a specific distribution. If the unsupervised ASD system can accurately model this distribution in its latent space, overall performance may improve.

Additionally, this study only evaluated the proposed method with sounds at an SNR = 6 dB, which involved minimal environmental noise. Since phase-based features are sensitive to noise, performance can suffer in low SNR conditions. Therefore, it is important to investigate how various environmental factors affect phase-based features in ASD.

Finally, exploring the combination of amplitude and phase information may facilitate more comprehensive detection. This approach should be addressed in future research to tackle more complex conditions in the field.

Publications

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

[2] Tran-Quang-Tuan Vo, Quoc-Huy Nguyen, and Masashi Unoki, "Anomalous Sound Detection Using Time-Frequency Derivative of Instantaneous Phase Features," accepted for publication in APSIPA ASC 2025.

Bibliography

[1] Y. Li, X. Li, Y. Zhang, M. Liu, and W. Wang, "Anomalous Sound Detection Using Deep Audio Representation and a BLSTM Network for Audio Surveillance of Roads," IEEE Access, vol. 6, pp. 58043–58055, 2018.

[2] 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.

[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, 2019.

[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009.

[5] Koizumi, Y. Kawaguchi, Y. Imoto, K. Nakamura, T. Nikaido, Y. Tanabe, R. Purohit, H. Suefusa, K. Endo, T. Yasuda, M. and Harada, N., "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.

[6] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep Learning for Anomaly Detection," ACM Computing Surveys, vol. 54, no. 2, pp. 1–38, 2021.

[7] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," Speech Commun., vol. 81, pp. 1–29, 2016.

[8] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and recent advances," in IEEE Signal Processing Magazine, vol. 32, no. 2, pp. 55-66, 2015.

[9] Kawaguchi, Y. Imoto, K. Koizumi, Y. Harada, N. Niizumi, D. Dohi, K. Tanabe, R. Purohit, H. and Endo, T., "Description and discussion on DCASE 2021 Challenge Task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, pp. 186–190, 2021.

[10] Dohi, K. Imoto, K. Harada, N. Niizumi, D. Koizumi, Y. Nishida, T. Purohit, H. Tanabe, R. Endo, T. Yamamoto, M. and Kawaguchi, Y., "Description and discussion on DCASE 2022 Challenge Task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in Proc. Detection and Classification of Acoustic Scenes and Events Workshop (DCASE Workshop), pp. 1–5, 2022.

[11] Dohi, K. Imoto, K. Harada, N. Niizumi, D. Koizumi, Y. Nishida, T. Purohit, H. Tanabe, R. Endo, T. and Kawaguchi, Y., "Description and discussion on DCASE 2023 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. 31–35, 2023.

[12] 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.

[13] 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.

[14] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, and Y. Kawaguchi, "MIMII DUE: sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA), pp 21–25, 2021.

[15] 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 Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, 2021.

[16] 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," in Proc. Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, 2022.

[17] A. Ito, A. Aiba, M. Ito and S. Makino, "Detection of Abnormal Sound Using Multi-stage GMM for Surveillance Microphone," in Proc. Int'l Conf. Info. Assurance and Security, pp. 733-736, 2009.

[18] C. F. Chan and W. M. Eric, "An abnormal sound detection and classification system for surveillance applications," in Proc. of Eur. Signal Process. Conf. (EUSIPCO), pp. 1851-1855, 2010.

[19] S. Zhao, "Acoustic anomaly detection based on similarity analysis," in Proc. Detection Classification Acoustic Scenes Events Challenge, pp. 1–3, 2020.

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

[21] H. Van Truong, N. C. Hieu, P. N. Giao, and N. X. Phong, "Unsupervised detection of anomalous sound for machine condition monitoring using fully connected U-Net," J. ICT Res. Appl., vol. 15, no. 1, pp. 4155, 2021.

[22] 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.

[23] 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 Challenge, pp. 1–5, 2022.

[24] 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.

[25] T. Hayashi, T. Komatsu, R. Kondo, T. Toda, and K. Takeda, "Anomalous sound event detection based on WaveNet," in Proc. of Eur. Signal Process. Conf. (EUSIPCO), pp. 2494–2498, 2018.

[26] B. Han, Z. Lv, A. Jiang, W. Huang, Z. Chen, Y. Deng, et al., "Exploring large scale pre-trained models for robust machine anomalous sound detection," in Proc. ICASSP, pp. 1326–1330, 2024.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in Proc. NIPS, vol. 33, pp. 12449–12460, 2020.

[28] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang, "Unispeech: Unified speech representation learning with labeled and unlabeled data," in Proc. ICML. PMLR, pp. 10937–10947, 2021.

[29] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Trans. ASLP., vol. 29, pp. 3451– 3460, 2021.

[30] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, "Autoencoders and their applications in machine learning: A survey," Artif. Intell. Rev., vol. 57, no. 2, pp. 1–19, 2024.

[31] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in Proc. ICLR, 2019.

[32] 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 2020 (DCASE) Workshop, pp. 46–50, 2020.

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

[34] 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.

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

[36] 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.

[37] R. D. Patterson and J. L. Holdsworth, "A functional model of neural activity patterns and auditory images," Advances in Speech, Hearing and Language Processing, (W. A. Ainsworth, ed.), Vol 3. JAI Press, London, 1991.

[38] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2.

[39] 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.

[40] K. Wilkinghoff and A. Cornaggia-Urrigshardt, "On choosing decision thresholds for anomalous sound detection in machine condition monitoring," in 24th International Congress on Acoustics (ICA), The Acoustical Society of Korea, 2022.

[41] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, Signals & systems (2nd ed.). USA: Prentice-Hall, Inc., 1996.

[42] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. II. Algorithms and applications," in Proceedings of the IEEE, vol. 80, no. 4, pp. 540-568, 1992.

[43] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," Lecture Notes in Computer Science, vol. 3644, pp. 878–887, 2005.

[44] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18–28, 1998.

[45] 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.

[46] 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.

[47] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, vol. 9, no. 11, 2008.

[48] 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.

[49] 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.

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