Improving Anomaly Detection Performance in Factory Sound Data through Noise Reduction

Jong Hyuk Lee
Electronic and Electrical Engineering
Kyungpook National University
Daegu, Korea
leewer354@knu.ac.kr

Min Young Kim
Electronic and Electrical Engineering
Kyungpook National University
Daegu, Korea
minykim@knu.ac.kr

***Abstract*—Noise in industrial environments degrades the performance of deep learning-based anomaly detection. This paper proposes a noise filtering technique to improve signal quality during the preprocessing stage, rather than increasing model complexity. It establishes a Global Noise Baseline and applies 3-segment masking to enhance signal quality. This preserves core signals even in complex noise, improving deep learning model accuracy by over 10%. Future applications include lightweight models and edge computing for deployment across diverse industrial settings.**

***Index Terms*—Noise filtering, Anomaly detection, Industry 4.0, time-series data**

I. Introduction

The Fourth Industrial Revolution and the shift to smart factories have elevated data-driven management and optimization of manufacturing processes to a core competitive advantage. Predictive maintenance, in particular, is gaining attention as a technology that minimizes production downtime due to equipment failures and reduces maintenance costs. Among these, acoustic-based diagnostics offer significant practicality as they can detect early-stage defects using only low-cost sensors [1-3]. However, in actual factories, signal quality degrades due to various machine noises and environmental sounds, leading to reduced performance in deep learning-based anomaly detection. Previous research has primarily pursued performance improvement through complex model structures, but this approach consumes significant computational resources, limiting its applicability in lightweight environments. Therefore, this paper proposes a novel noise filtering technique that enhances signal quality during the preprocessing stage. The proposed technique derives a stable noise baseline based on statistical analysis of the dataset and maximizes the signal-to-noise ratio (SNR) while preserving key signals through soft threshold-based 3-segment masking. Furthermore, it incorporates a logic filter specialized for specific high-energy periodic signal data, ensuring flexibility and safety.

II. Related Work

This section examines the core concepts of the proposed technique. SNR-Aware Masking employs deep learning to precisely estimate a signal's SNR, generating and applying a mask proportional to that value [4-5]. It preserves strong signal regions while gently suppressing noisy areas, dynamically adjusting noise intensity to match each data characteristic. This technique enables reliable signal separation in environments where noise and signals coexist. Conditional masking is a flexible filtering technique that dynamically changes the mask shape and application method based on predefined rules and conditions [6-7]. In this paper, we utilize multiple conditions to create a customized filtering strategy that suppresses, preserves, and enhances signals. This enables signal processing optimized for specific purposes, beyond simple noise suppression.

III. Experiment

This section describes the composition of the dataset used, the proposed noise removal method, the deep learning anomaly detection model, and the training and evaluation procedures. Figure 1 shows the overall framework of the proposed model.

A. Dataset

The data used in the experiment was the Malfunctioning Industrial Machine Investigation and Inspection [8]. The MIMII dataset is hierarchically organized based on real factory environment noise levels, machine types, and product models. The data consists of 10-second audio recordings categorized into normal and abnormal samples. All abnormal data is designated as test data. An equal number of corresponding normal data points are included in the test data, while all remaining normal data is used as training data. This ensures consistent and reproducible dataset partitioning.

B. Noise filter

Convert a single-channel audio file into a Mel Spectrogram using a frame size of 1024, a hop size of 512, and a Mel filter of 64. Five frames from the generated Mel Spectrogram are combined to form a two-dimensional input vector of dimensions 320x309.

First, we estimate a reliable reference value, the global noise floor, from the entire training dataset consisting of normal data. For the entire audio file consisting of normal data ($i = 1, \ldots, N$), we find the minimum energy value for each mel band ($l = 1, \ldots, n_{\text{mels}}$) across the entire time axis

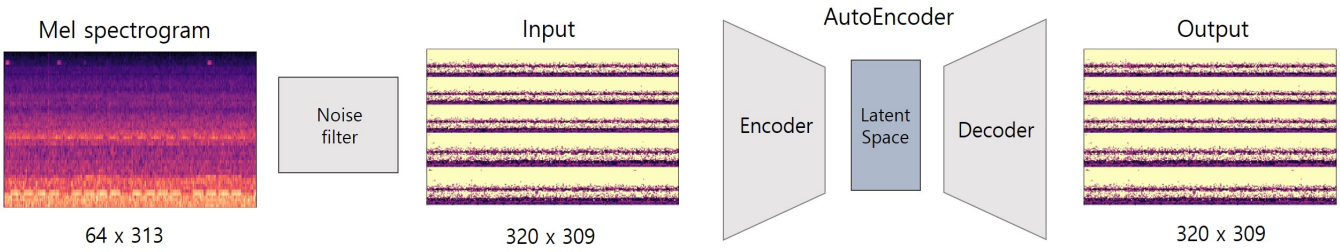

Fig. 1. The framework of the proposed model.

in each frequency band. This value becomes the "minimum noise profile."

$$v_{i,l} = \min_{t=1,\ldots,T_i} L_i(l,t)$$

The minimum energy profile (v_i) of each individual file is arithmetic averaged to obtain a single average noise profile ($\bar{\mathbf{v}}$) representing the entire dataset. This process leaves only the average noise characteristics.

$$\bar{\mathbf{v}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i$$

Finally, we select the top 10% of the highest-energy frequency bands from the average noise characteristics and set those values as the final global noise floor. In other words, we set the highest-energy frequency among the values classified as noise as the noise floor.

$$\theta_{\text{noise}} = \text{mean}(S_{\text{top}})$$

The average of these energy values is defined as the final global noise floor, θ_{noise}. This process produces a final noise floor that more accurately represents the actual background noise level, effectively treating all energy below this value as noise.

C. Conditional mask

A pre-calculated noise floor is used as a threshold to transform the energy values in the spectrogram. Sections where the energy value is significantly higher than the noise floor are considered key information and amplified by a factor of 2, helping subsequent models more easily capture important features. Conversely, energy below the noise floor is judged as pure background noise, and its value is set to 0 to completely remove unnecessary information. Furthermore, energy values near the noise floor are preserved unchanged. This prevents artificial artifacts caused by abrupt value changes at the threshold boundary and makes the overall transformation smoother. This transformation process maximizes the key signal and effectively removes noise, significantly improving data quality.

$L'(f,t)$ is log-Mel value after masking. δ_{Margin} is +5dB. This transformation process significantly improves data quality by maximizing core signals and effectively removing noise.

$$L'(f,t) = \begin{cases} 2 \cdot L(f,t), & \text{if } L(f,t) \geq \theta_{\text{noise}} + \delta_{\text{Margin}} \\ L(f,t), & \theta_{\text{noise}} \leq L(f,t) < \theta_{\text{noise}} + \delta_{\text{Margin}} \\ 0, & \text{if } L(f,t) < \theta_{\text{noise}} \end{cases}$$

D. Anomally detection model

An autoencoder (AE), an unsupervised neural network, is then trained to reconstruct the input as an anomaly detection model. The model is trained using the Adam optimizer and the Mean Squared Error (MSE) loss function, which measures the difference between the original input vector (X) and the reconstructed output ($\hat{x}$).

$$\min_{\theta,\phi} \mathbb{E}_{x\sim p_{\text{data}}}[\mathcal{L}(x, g_\phi(f_\theta(x)))]$$

The training and testing were performed on identical computers: an Intel i7-10700, an Nvidia Geforce RTX 2080, and a batch size of 512. The entire process is implemented in Python using the PyTorch framework. The model is trained for 50 epochs.

RESULT

Figure 2 shows the results after applying the noise filter. The original is the signal's mel-spectrogram image. The apply condition mask and SNR-swap mask show the result images when applying only one mask each. The proposed shows the result of the proposed noise filter. The system's performance is quantified using the Area Under the Receiver Operating Characteristic Curve (AUC). The reconstruction error for each vector is the mean squared error between the input and output, which determines whether the model is normal or abnormal based on how well it reconstructs the data. The anomaly score for an entire audio file is calculated as the average of the reconstruction errors for all vectors from that file. AUC is a comprehensive performance indicator that indicates how well the model distinguishes between normal and abnormal data, and takes a value between 0 and 1. A value closer to 1 indicates excellent performance, meaning the model perfectly distinguishes between normal and abnormal data. A value closer to 0.5 indicates poor model performance. A value closer

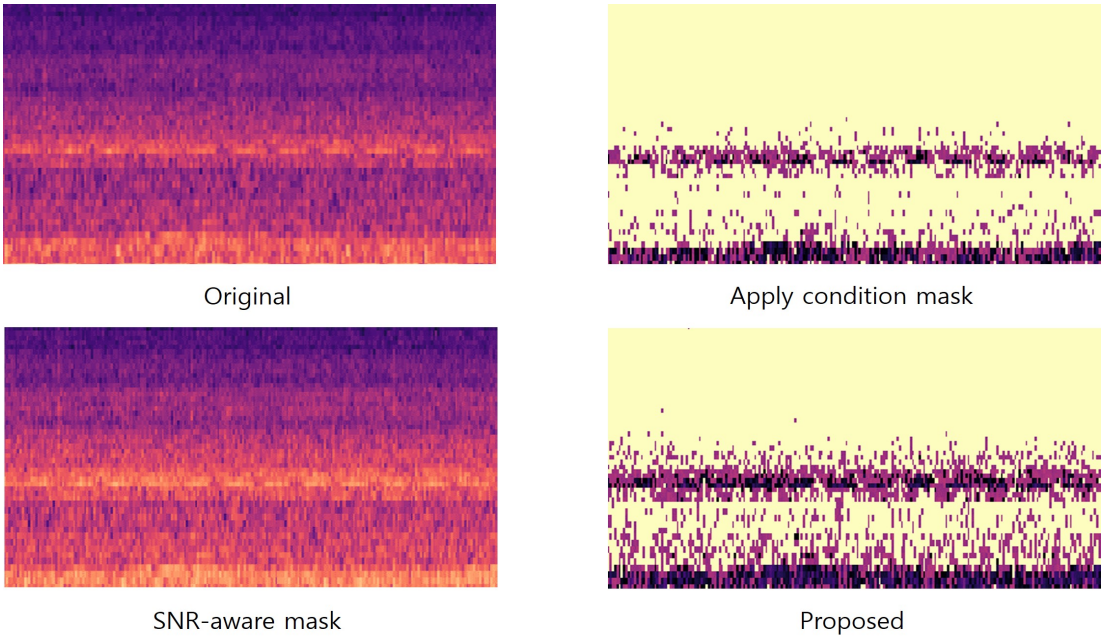

Fig. 2. The result image of noise filter.

to 0 indicates that the model is predicting the opposite of normal and abnormal data.

Table 1 summarizes the anomaly detection performance of the proposed noise filtering method compared with existing approaches. Overall, the proposed noise floor–based masking achieved the highest average performance of 0.68, surpassing the original baseline 0.58, the SNR-aware masking 0.58, and the conditional masking 0.66. Consistent improvements were observed across machine types. In particular, the proposed method significantly improved the performance of Pump, from 0.57 to 0.77, and Slider, from 0.61 to 0.72. Fan performance improved moderately, from 0.70 to 0.72, while Valve showed only a limited increase, from 0.46 to 0.49. Under various SNR conditions, the proposed method consistently improved detection performance, demonstrating notable robustness. This indicates that the method effectively handles noisy environments. In addition, performance gains were consistently observed across all device IDs. For example, the score for ID_00 improved from 0.54 to 0.67, and ID_06 from 0.54 to 0.67, highlighting the robustness of the proposed method against device-specific variations. In summary, the proposed masking consistently outperforms existing single masking approaches, and it is particularly effective in noisy environments and for machine types with complex acoustic characteristics such as pumps and sliders.

CONCLUTION

This study proposes a novel noise filtering and data preprocessing method that enhances the signal quality of machine acoustic data collected in industrial environments without increasing the computational load of deep learning models. This approach establishes a global noise baseline and applies soft threshold-based three-directional masking and adaptive masking to suppress background noise while preserving key signal components. This method focuses on improving feature quality before inputting data into the autoencoder model. It automatically learns a machine-specific background noise baseline using normal operating data, guiding a two-step adaptive filtering process that maximizes the signal-to-noise ratio. The core elements—global noise floor estimation and conditional masking—analyze the minimum energy values across multiple normal datasets to establish a stable baseline, suppress background noise, and emphasize the primary signal. This enables the data to self-identify signal-to-noise ratios and flexibly adapt to various machines and environments. Experimental results demonstrate that the proposed preprocessing method outperforms the unfiltered approach. This shows that

TABLE I
COMPARISON WITH EXISTING NOISE FILTERS BY TYPES

Type	Original	SNR-aware Mask	Apply conditional mask	Proposed
Total	0.58	0.58	0.66	**0.68**
Fan	0.70	0.65	0.69	**0.72**
Pump	0.57	0.59	0.75	**0.77**
Slider	0.61	0.64	**0.76**	0.72
Valve	0.46	0.45	0.43	**0.49**
0dB	0.59	0.59	0.67	**0.68**
6dB	0.63	0.62	**0.72**	0.71
min6dB	0.54	0.55	0.59	**0.63**
ID_00	0.54	0.58	0.60	**0.67**
ID_02	0.62	0.62	**0.70**	0.69
ID_04	0.63	0.59	0.63	**0.68**
ID_06	0.54	0.55	**0.70**	0.67

preprocessing alone can significantly enhance anomaly detection performance without modifying the model architecture.

Future research will explore dedicated noise models reflecting valve-specific operational and acoustic characteristics, along with strategies for selectively enhancing valve signals. This consideration stems from the relatively limited performance improvement observed in valve data.

ACKNOWLEDGMENT

This research was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE)(P0020536, The Competency Development Program for Industry Specialist), the Core Research Institute Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (RS-2021-NR060127) and the Regional Innovation System & Education (RISE) Glocal 30 program through the Daegu RISE Center, funded by the Ministry of Education (MOE) and the Daegu, Republic of Korea (2025-RISE-03-001).

REFERENCES

[1] J. H. Lee and M.Y. Kim, "Manufacturing Quality Management Based on TimeGAN and Seq2Seq ModelsWith Magnetic Press Machine Data," International Journal of Control, Automation and Systems, vol. 23, no. 4, pp. 1199-1209, 2025.

[2] J. H. Lee and M.Y. Kim, "Machine Learning-based Automatic Optical Inspection System with Multimodal Optical Image Fusion Network," International Journal of Control, Automation and Systems, vol. 19, no. 10, pp. 3503-3510, 2021.

[3] F. You, D. Wang, G. Li and C. Chen , "Fault diagnosis method of escalator step system based on vibration signal analysis," International Journal of Control, Automation and Systems, vol. 20, no. 10, pp. 3222-3232, 2022.

[4] S. W. Fu, Y. Tsao and X. Lu,, "SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement," In Interspeech, pp. 3768-3772, 2016.

[5] B. Kowalewski, T. Dau and T. May, "Perceptual evaluation of signal-to-noise-ratio-aware dynamic range compression in hearing aids," Trends in Hearing, vol. 24, 2020.

[6] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, A. S. Rif, J. W. Ron, Y. Jia and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," arXiv preprint arXiv, 2018.

[7] E. Cano, J. Nowak and S. Grollmisch, "Exploring sound source separation for acoustic condition monitoring in industrial scenarios," In 2017 25th European Signal Processing Conference (EUSIPCO), pp. 2264-2268, 2017.

[8] 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," arXiv preprint arXiv, 2019.