# Acoustic Monitoring in Industrial Plants with Autoencoders and a Mobile Robot

Hayato Fujita[1], Jun Younes Louhi Kasahara[1], Shinji Kanda[1], Keiji Nagatani[1], Seiji Kasahara[2], Seigo Fukumoto[2], Sunao Tamura[2], Toshiya Kato[2], Masahiro Korenaga[2], Akinobu Sasamura[2], Misaki Hoshi[2], Hajime Asama[1] and Atsushi Yamashita[1]

***Abstract*— Industrial Plants such as refineries are complex installations that require constant monitoring and inspection to ensure safe and stable operation. This is currently conducted by field operators and one of the important tasks is acoustic inspection, i.e., listening for abnormal sounds while on patrol, skills of which are difficult to describe in operation procedures. Due to the issues of skilled staff retirement, costs, and inspection quality variance, the automation of acoustic inspection is desirable. Due to the large scale of these installations, several acoustic landscapes co-exist. This makes the establishment of a single model for abnormal sound detection difficult. Therefore, considering a mobile robot patrolling the plant, this study proposes to divide the robot's path into a grid where in each grid cell a distinct model is trained, bypassing the issue of differing acoustic landscapes. Experiments conducted in a simulated environment confirmed the effectiveness of the proposed method.**

## I. Inspection of Industrial Plants

While the recent trend is towards the reduction of carbon emissions due to ecological concerns, industries related to the processing of crude oil remain a major pillar of modern societies by providing critical products such as gasoline for automobiles and plastic goods. At the core of the oil processing industry are complex plants such as the one pictured in Fig. 1.

Fig. 1: Kawasaki Refinery (ENEOS Corporation)

In such installations, one major concern is the maintenance of the machinery: abnormal operation due to aging or malfunctions may lead to costly production losses and/or expensive repairs if left unchecked for a long time. Therefore, plants usually have dedicated groups of field operators who perform inspection patrol on a high frequency, up to several rounds per day. Such patrols include conducting a comprehensive inspection of the equipment while walking around the plant. In Table I are reported some common anomalies that can be detected by human senses during their usual field patrol. It can be seen that field operators apply their sight, hearing, smell and feeling temperature to detect anomalies in practice. One issue is that individual differences between field operators directly affect the quality of the conducted inspection. Additionally, the process is laborious and manpower heavy, resulting in high costs. Therefore, the automation of plant inspection patrol is highly desirable.

Mainly around rotating machinery such as pumps and compressors, sound is an important factor to detect anomalies. For example, dry/scratched bearing can be conveniently detected during operation by the high pitch noise they produce. However it is challenging since several machines operate in close proximity of each another in a plant and thus the environment is noisy.

Towards this goal, the development of a mobile robot for plant inspection is ongoing [1]. Such mobile robots can move on flat grounds as well as stairs due to their crawlers: they can therefore follow the same inspection patrol path as field operators. Due to the potential presence of combustible gases in the environment, sensors of such robots need to adequately protected to prevent possible ignition of the surrounding atmosphere. This raises the costs of sensor installation, thus the available number of sensors is limited.

Audio-based anomaly detection can be tackled from several perspectives. A classification approach using supervised learning, i.e., gathering training data of normal and abnormal sound and training a classifier model, has been traditionally a popular one. Convolutional Neural Networks are often employed in this task, such as in [2], [3] or [4]. While highly performing classifiers are reported, it is conditioned by the availability of adequate training data. Normal sound data can be easily obtained from normal operation of the target machinery. However, anomalies are rare in comparison, thus gathering enough abnormal sound data for training is difficult. Additionally, anomalies correspond to any non-normal operation, therefore it is impossible to collect data

[1]The University of Tokyo, Tokyo 113-8656, JAPAN, fujita@robot.t.u-tokyo.ac.jp

[2]ENEOS Corporation, Tokyo 100-8162, JAPAN

TABLE I: Common issues field operators look for during their usual inspection patrol (obtained through interviews with actual workers in the field, ENEOS Corporation).

| **Machinery type** | **Common issues** | **Detection means** |
|---|---|---|
| Pump | Bearing abnormal sound | Hearing |
| | White smoke caused by leakage | Sight and smell |
| | Ignition due to leakage | Sight and smell |
| | Low oil level | Sight |
| | Decrease/loss of coolant due to clogging, etc. | Touch |
| | Abnormal heat emission | Feeling temperature |
| Pipe | Surface corrosion | Sight |
| | Aging of thermal insulation | Sight |
| | Abnormal surface temperature | Feeling temperature |
| | Leaks due to corrosion | Sight and smell |
| | Gas/oil leakage at flanges | Sight and smell |
| Gauge | Abnormal values | Sight |
| Miscellaneous | Heat generation from furnace wall due to fallen refractory bricks | Sight |
| | Inadequate valve opening/closing | Sight |

for all possible classes of anomaly.

In that sense, unsupervised learning approaches could be considered more suited for anomaly detection. This term has historically loosely regrouped various approaches that did not fit the definition of supervised learning. In recent years, Autoencoders, consisting of learning a *normal model* using only normal data to then discriminate any outlier data as abnormal, have gained high popularity and seen successful application to abnormal sound detection [5][6]. However, those were limited to clean and cured datasets or single machine monitoring. Plants are large sites that often span over several square kilometers, with possibly hundreds, if not thousands, of unique sound sources due to the made-to-order nature of most machinery of the field and the spatially irregular arrangement of these machines. This results in several *acoustic landscapes*, i.e., environments with differing sounds, coexisting in a single large environment. Establishing a single normal model for such an environment is therefore difficult: the normal sound in one location differs greatly from the one in another location within the same plant.

Therefore, the objective of this paper is to achieve abnormal sound detection for a mobile robot in environment containing several acoustic landscapes.

## II. Acoustic Defect Detection

As previously stated, the environment considered in this study is one containing several acoustic landscapes with differing characteristics. Two states are defined: normal and abnormal. The normal state corresponds to all equipment functioning normally. Therefore, all pieces of equipment emit their *normal sound*. When a piece of equipment fails, it emits a different sound, which we designate as *abnormal sound*. This is the abnormal state. While normal sound data is available for use in training, abnormal data is not. A mobile robot patrols along a route near each piece of equipment. Using a microphone mounted on the robot, the goal is to accurately detect the abnormal sound of a piece of equipment when the robot is in its vicinity.

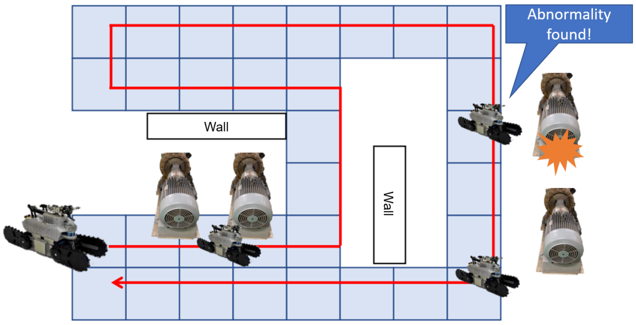

Fig. 2: Concept of the proposed method: the patrol route of the robot is divided into a grid and an Autoencoder is trained on each grid.

### A. Concept

The challenge in the considered environment is the coexistence of various acoustic landscapes. A normal model for the whole plant, if possible, would therefore be highly complex, require large amounts of training data, as well as long training time and resources. Therefore, the concept of the proposed method is to divide the environment into smaller sub-environments: this would result in each sub-environment being limited to only a single acoustic landscape. Concretely, this is achieved through division of the patrol route of the robot into a grid. Afterwards, in each grid cell is trained an Autoencoder using only the training data collected on that particular grid cell. During the actual patrol, the Autoencoder of the specific grid cell on which the robot is currently located is used to perform the inspection. This is illustrated in Fig. 2.

### B. Training data acquisition and pre-processing

In each grid cell, the sound signal is gathered through a microphone. The sound signal, initially a time-series data, is transformed into Fourier spectrum using the Short-time Fourier Transform with a sliding Hanning window with overlap. Those would serve as inputs to the Autoencoder.

### C. Autoencoder Architecture

The Autoencoder is one of the unsupervised Deep Learning methods and can be used for anomaly detection by training only on normal data. The Autoencoder is comprised of two symmetrical components: the encoder and the decoder. The encoder has weights $\mathbf{w}_\text{e}$ and biases $\mathbf{b}_\text{e}$ as parameters, as in Eqs. (1,2). Similarly, the decoder has weights $\mathbf{w}_\text{d}$ and biases $\mathbf{b}_\text{d}$ as parameters, as in Eqs. (3,4). During the training process, samples are inputted into the encoder first and reduced into a lower dimension feature space, often referred to as the latent space in the literature. Following that, the decoder reconstructs the sample into its original dimension. For input $\mathbf{A} = (a_1, ..., a_N)$ and output $\mathbf{B} = (b_1, ..., b_N)$ of the Autoencoder, the Mean-Square Error Loss (MSELoss) is defined as in Eq. (5) and training is conducted so as to minimize it. Therefore, the Autoencoder learns to compress and decompress the data with minimal loss of information, i.e., without reconstruction error.

$$\mathbf{u}_\text{e} = \mathbf{w}_\text{e} \cdot \mathbf{x}_\text{e} + \mathbf{b}_\text{e} \tag{1}$$

$$\mathbf{x}_\text{e} = \text{relu}(\mathbf{u}_\text{e}) \tag{2}$$

$$\mathbf{u}_\text{d} = \mathbf{w}_\text{d} \cdot \mathbf{x}_\text{d} + \mathbf{b}_\text{d} \tag{3}$$

$$\mathbf{x}_\text{d} = \text{relu}(\mathbf{u}_\text{d}) \tag{4}$$

$$\text{MSELoss}(\mathbf{A}, \mathbf{B}) = \sqrt{(a_1 - b_1)^2 + ... + (a_N - b_N)^2} \tag{5}$$

### D. Threshold Value Selection

During inference, a threshold is used on the previously mentioned reconstruction error to assess whether the considered sample belongs to the same data distribution as the ensemble of data provided during training, i.e., if it belongs to the normal sound data distribution of not. If a normal sound sample is considered, the reconstruction error should be low since the Autoencoder was trained on such data. On the other hand, if a abnormal sound sample is considered, the differing Spectrum characteristics should not allow the Autoencoder to successfully compress and decompress the sample and the reconstruction error should be large.

To determine the appropriate value for the threshold $T$, the reconstruction error of the Autoencoder on the training data $(\mathbf{X}_1, ..., \mathbf{X}_{N_\text{s}})$ after training is approximated as a continuous probability density distribution $p$ by Kernel Density Estimation (KDE) using kernel $K$ as in Eq. (6), where $h$ is the bandwidth. The value for the bandwidth $h$ is selected as in Eq. (7) following [7], with IQR the interquartile range and $\sigma$ the standard deviation of the kernel function. This enables setting the threshold which would consider 90% of the training dataset as normal data.

$$p(x) = \frac{1}{N_\text{s} * h} \sum_{i=1}^{N_\text{s}} K(\frac{x - \mathbf{X}_\text{i}}{h}) \tag{6}$$

$$h = 0.9 \text{min}(\sigma, \frac{\text{IQR}}{1.34}) \tag{7}$$

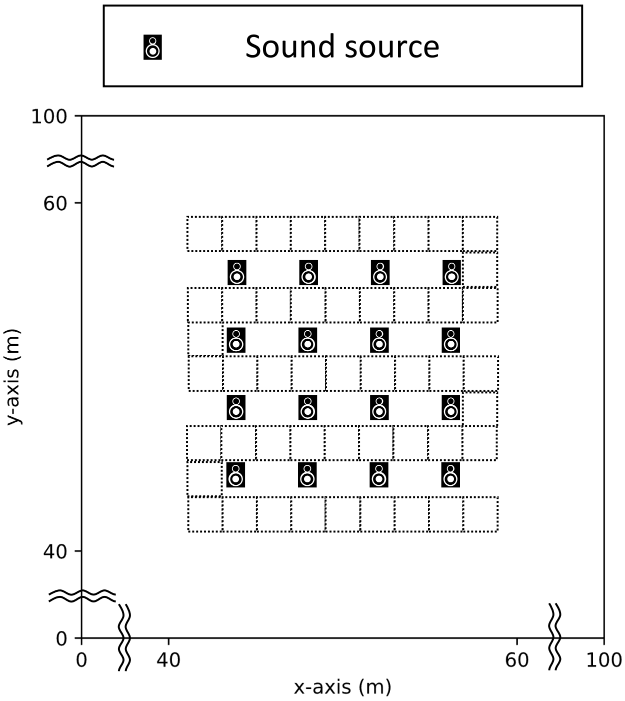

Fig. 3: Schematic of the environment built using Pyroomacoustics and MIMII dataset.

## III. Experiments

### A. Setup

Experiments were conducted through simulations to confirm the effectiveness of the proposed method. Using Pyroomacoustics [8], a square room containing 16 sound sources in a 4 by 4 configuration was established. For each sound source, the sound of a distinct machine contained in the MIMII dataset [9] was used. The schematic of the simulation environment is provided in Fig. 3. The path of the mobile robot was considered as a serpentine one going between the rows of sound sources and was divided into a grid with each grid cell being 2 m by 2 m. This resulted in 50 grid cells for the considered environment. In each grid cell was collected 5168 training sample data. The duration of each training data sample was about 0.2 s at sampling rate 22050 Hz. Those were converted into Fourier spectrum with Hanning window of size 4096. In a similar fashion, test data containing 1292 normal samples and 1292 abnormal samples over the whole environment was collected.

The considered Autoencoders were all limited to 4 layers deep, of dimensions 2049, 512, 128 and 32. The learning rate and weight decay were manually set at 0.08439 and 0.00078, respectively.

Two scenarios were considered: A, were all sound sources were normal, and B, where one sound source was abnormal with the remaining 15 being normal.

### B. Results and Discussions

The results obtained using a single Autoencoder for the whole environment and those obtained using the proposed

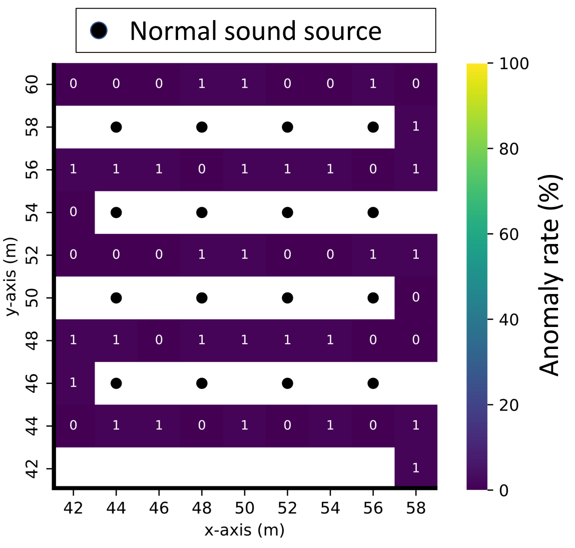

Fig. 4: Result of a single Autoencoder in scenario A.

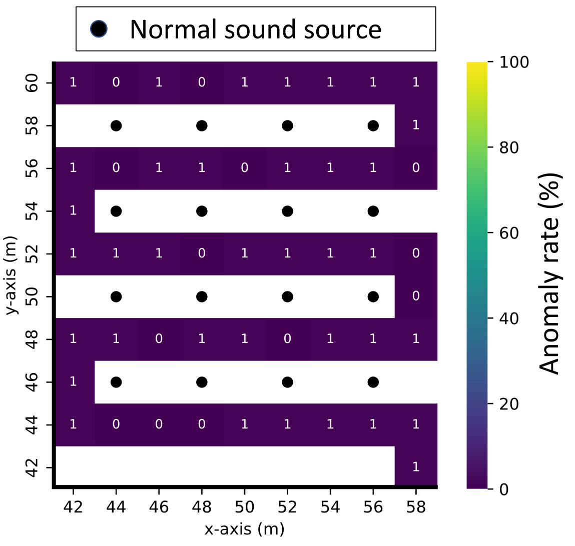

Fig. 5: Result of the proposed method in scenario A.

method are shown in Figs. 4 and 5, for scenarios A and B, respectively. For each grid cell is reported the anomaly rate, i.e., the percentage of the test data classified as abnormal by each method. Therefore, the ground truth for the anomaly rate is 0% for scenario A and 100% for scenario B.

For scenario A, simulating a plant working normally, i.e., without any abnormal sound present, it can be seen that both methods achieved similar outputs. For each grid cell where evaluation was conducted, the anomaly rates were very low, under 1%. For scenario B, simulating a plant with a machine experiencing an issue, i.e., emitting abnormal sounds, differing outputs were observed for the considered methods. In the case of a single Autoencoder, the grid cells immediately adjacent to the source of the abnormal sounds achieved anomaly detection rates between 57% and 61%. However, for grid cells located further away from the

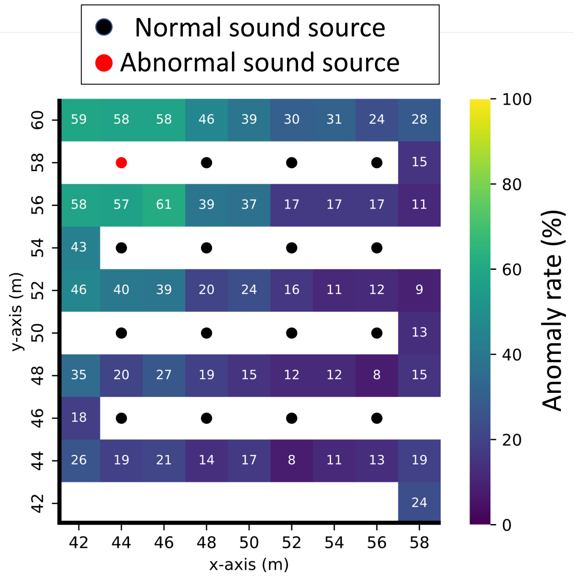

Fig. 6: Result of a single Autoencoder in scenario B.

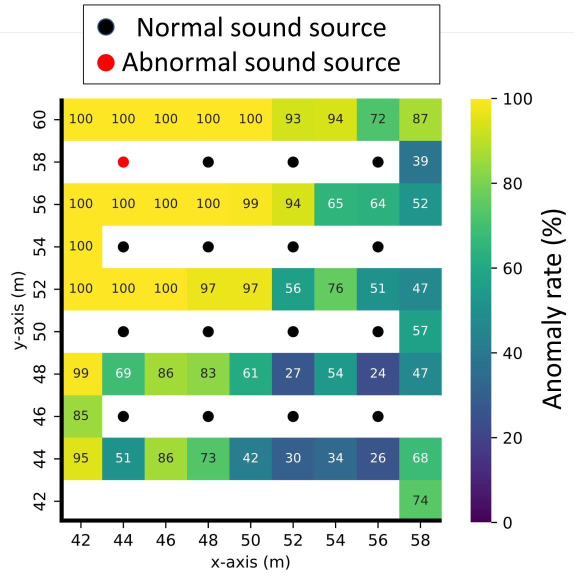

Fig. 7: Result of the proposed method in scenario B.

abnormal sound source, detection rates fell down to about 10%. This is due to two reasons. Firstly, as sound propagates in the environment, it attenuates, rendering detection more difficult at longer ranges. Secondly, the presence of normal sources results in a mix of abnormal and normal sounds being captured by the microphone. For grid cells further away from the abnormal sound source, the sound signal is predominantly composed of normal sounds, requiring a more sensitive classifier for detection. In comparison, the proposed method achieved much higher performance in scenario B. anomaly detection rates of 100% were achieved for several grid cells in the vicinity of the abnormal sound source. Additionally, high anomaly detection rates were obtained for grid cells further away from the abnormal sound source, about the double compared to the method with a single Autoencoder. This shows that the proposed method allows for higher detection rate of abnormal sounds as well as detection at longer distances.

## IV. Conclusion

In this paper was proposed a method for abnormal sound detection for a mobile robot in environments containing several acoustic landscapes. The proposed method bypassed the issue of establishing a model encompassing such a complex environment by dividing the path of the robot into a grid and training a normal sound model using an Autoencoder in each grid cell.

One issue of the proposed method is the number of models, which corresponds to the number of grid cells. Therefore, as the robot path increases, the number of models also increases. In the future, we plan to improve this approach by considering merging grid cells so that the number of models more closely match the number of acoustic landscapes. Additionally, we plan to conduct experiments in more realistic environments such as actual plants: in such settings, factors such as the influence of wind or sounds produced by the robot itself will need to be considered.

## References

[1] K. Shukutani, K. Onishi, N. Onishi, H. Okazaki, H. Kojima, and S. Kobori, "Development of explosion-proof autonomous plant operation robot for petrochemical plants," in *Mitsubishi Heavy Industries Technical Review*, vol. 55, no. 4, 2018, pp. 1–6.

[2] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "CNN architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[3] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 421–425.

[4] Y. Yao, H. Wang, S. Li, Z. Liu, G. Gui, Y. Dan, and J. Hu, "End-to-end convolutional neural network model for gear fault diagnosis based on sound signals," in *Applied Sciences*, vol. 8, no. 9. MDPI, 2018, p. 1584.

[5] D. Y. Oh and I. D. Yun, "Residual error based anomaly detection using auto-encoder in SMD machine sound," in *Sensors*, vol. 18, no. 5. MDPI, 2018, p. 1308.

[6] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 271–275.

[7] B. W. Silverman, *Density estimation for statistics and data analysis*. Routledge, 2018.

[8] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 351–355.

[9] 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 *arXiv preprint arXiv:1909.09347*, 2019.