DOI: 10.61782/fa.2025.0381

FORUM ACUSTICUM
EURONOISE 2025

EXPLOITING SPATIAL INFORMATION FOR ANOMALY DETECTION IN INDUSTRIAL MACHINES USING AUTOENCODERS

Clara Luzón-Álvarez[1*] **Ana M. Torres-Aranda**[2] **Francesc J. Ferri**[1]
Maximo Cobos[1]
[1] Department d'Informàtica, Universitat de València, Spain
[2] Departamento de Ingeniería Eléctrica, Electrónica, Automática y de Comunicaciones, Universidad de Castilla-La Mancha, Spain

ABSTRACT

Detecting machine failures or anomalies using sound remains a challenging task. In real-world environments, recording machine anomalies or failures is difficult, as they do not happen so often, limiting the systems to training on normal operational sounds. Additionally, the variability in environmental and machine conditions, such as speed, temperature, and background noise, further complicates the task. While significant progress has been made in recent years, much of the research has focused on mono audio processing. To explore whether multichannel audio can enhance model performance, we propose a modification of the DCASE Task 2 baseline model to support multichannel processing. Instead of processing all channels uniformly, our approach involves using one channel as a reference and calculating its difference from the others. Although each channel has its encoder and decoder, the embedded space is shared and passed to each decoder. The performance of this model is compared with the baseline, demonstrating slightly better results.

Keywords: *anomalous sound detection, autoencoder, spatial audio.*

**Corresponding author*: *clara.luzon@uv.es.*

1. INTRODUCTION

The goal of anomalous sound detection (ASD) is to determine whether a sound corresponds to normal operation or indicates an anomaly. The goal of anomalous sound detection (ASD) is to determine whether a sound corresponds to normal operation or indicates an anomaly. This task has various applications, such as in medical analysis [1], traffic control [2], and industrial monitoring and conditioning systems [3]. In the context of industrial machines, ASD aims to detect whether a machine is functioning normally or exhibiting an anomaly. However, collecting real-world anomalous sound data is challenging due to its rarity and high variability [4]. To address this, modern ASD methods typically train models on normal sound data, learning its distribution to distinguish normal from abnormal sounds.

A widely used approach involves autoencoders (AE), which learn compressed representations by reconstructing input data. The DCASE Task 2 baseline [3] employs a simple autoencoder trained on normal data, using reconstruction error as an anomaly score. Similarly, [5] applies an autoencoder for condition monitoring of rotating machines. More advanced methods integrate LSTM layers to improve detection [6], while others explore convolutional [7], hybrid [8], and variational autoencoders [9]. A GAN-based adversarial training approach is proposed in [10].

While effective, these approaches focus on monaural processing, using only single-channel audio and overlooking spatial characteristics. These have proven beneficial in tasks like sound source localization, speech enhancement, and acoustic scene analysis. For instance, in sound

source localization, [11] employs a U-Net-based model to extract spatial features from beamforming maps. In speech enhancement, [12] proposes a binaural method using a CNN-transformer architecture to improve intelligibility while preserving spatial cues. For acoustic scene analysis, [13] introduces the spatial cepstrum method to enhance robustness without requiring microphone synchronization.

Despite their success in other areas, spatial features remain underexplored in ASD. In this work, we propose a multi-channel autoencoder-based method that incorporates spatial information, analyzing audio using both single-channel data and inter-channel differences.

2. PROPOSED METHOD

Our proposal is based on extending the baseline model introduced in the DCASE Task 2 (2023) [3]. This model employs an autoencoder trained exclusively on normal sound data, using the reconstruction error as the anomaly score. To define a decision threshold, the authors assume that anomaly scores for normal sounds follow a gamma distribution. The parameters of this distribution are estimated based on the reconstruction errors from normal sound data in the training set for each machine. The threshold is then determined as the 90th percentile of the gamma distribution, with any score surpassing this value identified as anomalous. For model input, 128 bands of log-Mel energies are extracted using STFT with 64 ms frames and a 50% hop size. Five consecutive frames are concatenated to form input vectors for the autoencoder, resulting in a 640-dimensional representation. Figure 1(a) provides an example, where each column corresponds to one such input vector.

The model's architecture is depicted in Figure 2. From the audio waveform, input feature vectors (x^-) are computed as described earlier. Each vector is then passed through the encoder (Enc), which compresses it into an embedded space representation (z). The decoder (Dec) subsequently reconstructs the signal ($\hat{x}^-$) from this latent representation. For training, the AE model parameters θ are optimized to minimize the mean squared error (MSE) between a normal input sample x^- and its reconstructed output $\hat{x}^-$. This error also serves as the anomaly score during inference.

$$Loss = \text{MSE}(x^-, \hat{x}^-), \tag{1}$$

where

$$\hat{x}^- = \text{Dec}_\theta(z) = \text{Dec}_\theta(\text{Enc}_\theta(x^-)). \tag{2}$$

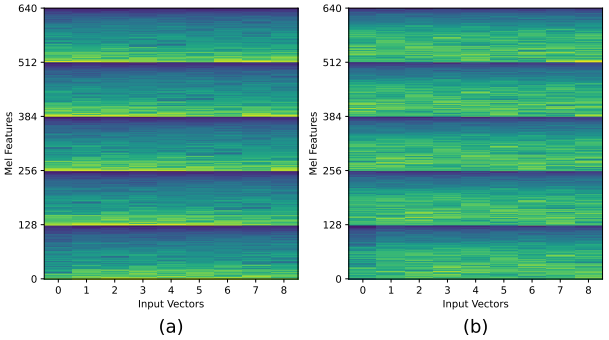

Figure 1. (a) Eight input vectors, each comprising five concatenated frames of the Mel spectrogram of the input channel on the baseline and the reference channel on our proposal. (b) Eight input vectors, each comprising five concatenated frames of the Mel spectrogram of the difference input channel.

The baseline model also included an alternative mode for computing the anomaly score. While the model architecture remains unchanged, this version replaces the standard mean squared error (MSE) with Mahalanobis distance as the anomaly score. During training, the model computes the residuals, the differences between normal inputs and their reconstructions, and uses these residuals to estimate a covariance matrix. This matrix models the expected distribution of reconstruction errors under normal conditions. After the final training epoch, this covariance matrix is fixed and used during inference to compute the Mahalanobis distance between the residual (the difference between the input x and its reconstruction $\hat{x}$) and the learned distribution.

$$\mathcal{A}_\theta(x) = \text{Mahalanobis}(x, \hat{x}, \Sigma^{-1}), \tag{3}$$

where Σ is the covariance matrix.

Several modifications are proposed to adapt this model for spatial processing. First, defining the model's input is a non-trivial task, as it is crucial to maximize feature learning while minimizing the size of the model. Our approach utilizes two channels recorded from opposite-side microphones: a reference channel (x), representing the signal from one microphone, and a difference channel (x'), computed as the signal difference between the two microphones. Both channels are processed similarly to the baseline, where 128 Mel features are extracted, and five consecutive frames are concatenated to form input vec-

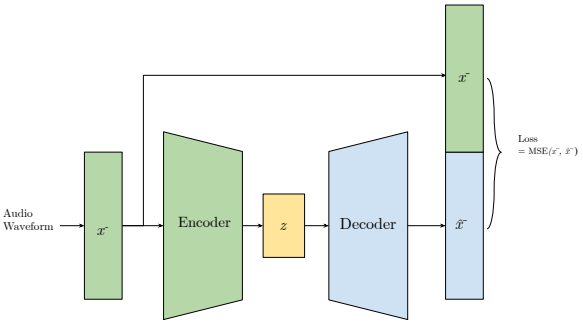

Figure 2. Baseline model from the DCASE challenge 2024 task [3]

tors. An example of these input vectors can be seen in Figure 1 where (a) corresponds to the reference channel and the (b) to the difference channel.

The model structure is illustrated in Figure 3. Each channel has its encoder and decoder. The input vectors (x^- and x'^-) are fed into their respective encoders, which generate the corresponding embedded space representations (z and z'). These representations are then concatenated and passed to both decoders, which attempt to reconstruct their respective input vectors.

To define the model's loss function, we combine two reconstruction errors:

$$\mathcal{L} = \mathcal{L}_{ref} + \alpha \mathcal{L}_{dif}, \tag{4}$$

where $\mathcal{L}_{ref}$ is the reconstruction error of the reference channel, and $\mathcal{L}_{dif}$ is the reconstruction error of the difference channel:

$$\mathcal{L}_{ref} = \text{MSE}(x^-, \hat{x}^-) \tag{5}$$

$$\mathcal{L}_{dif} = \text{MSE}(x'^-, \hat{x}'^-). \tag{6}$$

We introduce the hyperparameter α to balance the contribution of each channel's reconstruction error during training. This parameter controls the relative weighting of the difference channel's loss, allowing us to adjust its influence on the overall optimization. By tuning α, we can assess the impact of each channel on the model's performance and determine which contributes more effectively to improving anomaly detection.

Like the baseline, we extend our approach to the Mahalanobis distance-based anomaly scoring. We compute two covariance matrices: one for the reference channel

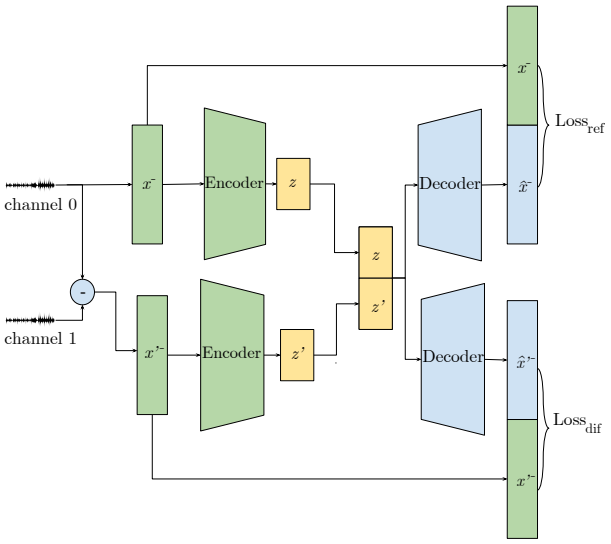

Figure 3. Proposed two-microphone model.

and one for the difference channel. From each one, we compute an anomaly score and combine them as:

$$\mathcal{A}_\theta(x) = \mathcal{A}_{ref}(x) + \alpha \mathcal{A}_{dif}(x'), \tag{7}$$

where

$$\mathcal{A}_{ref}(x) = \text{Mahalanobis}(x, \hat{x}, \Sigma_{ref}^{-1}) \tag{8}$$

$$\mathcal{A}_{dif}(x') = \text{Mahalanobis}(x', \hat{x}', \Sigma_{dif}^{-1}). \tag{9}$$

Here, Σ_{ref} and Σ_{dif} are the covariance matrices for the reference and difference channels, respectively.

3. EXPERIMENTS AND RESULTS

3.1 Dataset

We evaluate our method using the MIMII dataset [14], which includes recordings of four machine types: fans, slide rails, valves, and pumps. For each type of machine, recordings were made from multiple individual units. The sounds were captured as 16-bit audio signals sampled at 16 kHz. As shown in Fig. 4, recordings were made using a circular microphone array with eight microphones placed 50 cm from the machines, except for valves, which were recorded at 10 cm. Each session captured a single machine along with background noise.

For training, we randomly selected 1,000 normal audio samples, mixing different individual units and noise

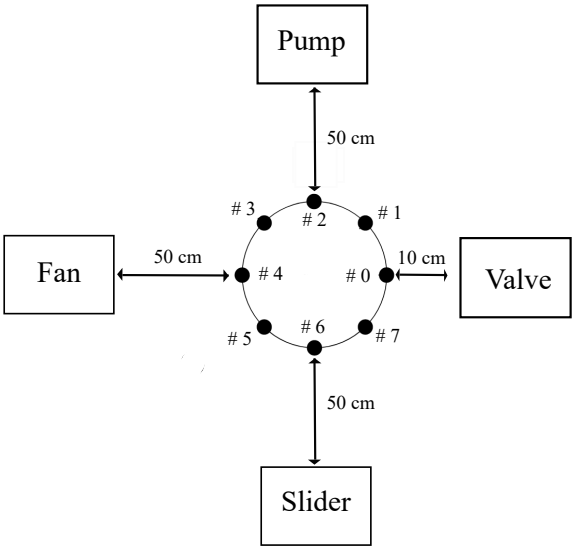

Figure 4. Recording scheme of the dataset (reproduced from [14])

levels. The test set included 200 normal and anomalous audio files, simulating conditions similar to those in the DCASE challenge.

3.2 Evaluation Metrics

To measure the performance of our proposal, we have used the same metrics used in the DCASE Challenge and by the majority of related works [4, 7, 9, 10]. We use the area under the ROC curve (AUC) and the partial AUC (pAUC). These metrics are computed as follows:

$$\text{AUC} = \frac{1}{N_- N_+} \sum_{i=1}^{N_-} \sum_{j=1}^{N_+} \mathcal{H}\left(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)\right), \quad (10)$$

$$\text{pAUC} = \frac{1}{\lfloor pN_- \rfloor N_+} \sum_{i=1}^{\lfloor pN_- \rfloor} \sum_{j=1}^{N_+} \mathcal{H}\left(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)\right), \quad (11)$$

where N_- and N_+ are the number of normal test samples (x^+) and the number of anomalous test samples (x^-), respectively. The pAUC is computed by defining a false-positive rate (FPP) p (in our case $p = 0.1$) and $\mathcal{H}(\cdot)$ denotes the Heaviside step function, returning 1 if its argument is positive and 0 otherwise.

3.3 Implementation

Following the same approach as the baseline, our proposed model is trained independently for each machine type, meaning a distinct model is optimized using data specific to that machine category. For each machine type, we train both our model and the baseline model using 1,000 raw waveform signals. For the log-Mel spectrogram computation, we use a frame size of 1,024 samples with 50% overlap, and apply a Mel filter bank consisting of 128 filters. The Adam optimizer is used for training. [15] to train the model with a learning rate of 0.001. The model is trained for 50 epochs, and the batch size is 256.

3.4 Performance comparison

We evaluated our model using different microphone pairs and α values for each machine. Table 1 summarizes the results across all configurations and scoring methods: Mean Squared Error (MSE) and Mahalanobis Distance (Mahala). The best-performing result for each machine and scoring method is highlighted in bold. Our model consistently outperforms the baseline across all machines and scoring metrics, demonstrating that optimizing the α parameter and selecting appropriate microphone pairs significantly enhances anomaly detection performance. Examining the average results, our method consistently surpasses the baseline, with MSE scoring showing the most significant improvement—AUC and pAUC increasing by 4.29% and 3.45%, respectively.

In Fig. 5, a detailed representation of the AUC (%) results is provided. Each subgraph shows the AUC performance for a specific machine, comparing two microphone pairs and analyzing the impact of the parameter α. The influence of microphone placement is clear, as the optimal pair varies by machine, emphasizing the need for proper setup selection. Regarding α, for both microphone combinations, at least one value always surpasses the baseline, showing that tuning this parameter can significantly improve performance.

4. CONCLUSIONS

In this paper, we have presented a self-supervised anomalous sound detection model that leverages autoencoders with spatial information. The proposed method demonstrates improved performance compared to the traditional single-channel audio processing model. This experiment opens the door for further exploration of spatial processing in other common ASD models and provides oppor-

Table 1. Results of all the experiments carried out with two different pairs of microphones and for scores computed using MSE and for Mahalanobis distances. The baseline results are included for comparison.

Machine	Metric	MSE									MAHALANOBIS								
		MICS 0-4				MICS 2-6				Baseline	MICS 0-4				MICS 2-6				Baseline
		α = 0,1	α = 0,5	α = 2,0	α = 10,0	α = 0,1	α = 0,5	α = 2,0	α = 10,0		α = 0,1	α = 0,5	α = 2,0	α = 10,0	α = 0,1	α = 0,5	α = 2,0	α = 10,0	
Fan	AUC	87,14	91,21	93,03	**94,03**	88,99	88,44	86,10	85,37	86,98	86,12	90,90	93,96	**94,46**	89,77	89,81	88,96	88,29	87,19
	pAUC	70,07	77,12	82,71	**85,78**	75,26	72,59	70,43	67,43	69,59	67,47	74,49	80,92	**83,48**	74,01	71,97	70,91	67,65	70,32
Slider	AUC	95,21	94,85	94,89	93,14	92,98	94,13	95,52	**96,00**	92,24	95,90	95,07	**95,46**	93,05	89,99	91,36	93,79	94,69	92,01
	pAUC	85,70	85,42	85,71	82,24	82,84	84,31	87,00	**87,89**	79,27	82,40	81,80	**82,21**	76,54	75,63	76,76	79,48	82,58	73,56
Pump	AUC	71,90	73,64	68,63	67,66	73,54	75,19	**75,79**	73,83	69,10	77,23	79,10	73,79	72,69	80,78	**82,88**	82,56	81,19	75,61
	pAUC	58,88	61,11	56,76	56,80	64,07	65,68	**65,57**	63,23	58,00	53,80	54,02	53,14	53,36	56,76	**58,63**	56,98	58,59	52,01
Valve	AUC	**58,24**	58,17	57,28	58,12	56,79	54,58	51,65	52,17	55,13	48,08	48,35	47,32	47,24	**49,15**	47,33	47,06	45,22	46,63
	pAUC	**48,68**	48,54	48,36	48,25	48,54	48,28	48,43	48,54	47,99	48,83	48,87	48,65	48,72	**48,36**	48,68	48,65	48,90	48,79
Average	AUC	75,39	**76,51**	75,02	74,97	75,19	74,70	73,21	72,93	72,80	71,70	**72,93**	71,68	71,10	72,76	72,37	72,38	71,02	70,17
	pAUC	63,03	**64,89**	64,46	64,36	64,97	64,96	64,94	63,87	61,48	60,58	**61,90**	62,60	62,20	61,51	62,00	61,74	62,10	59,23

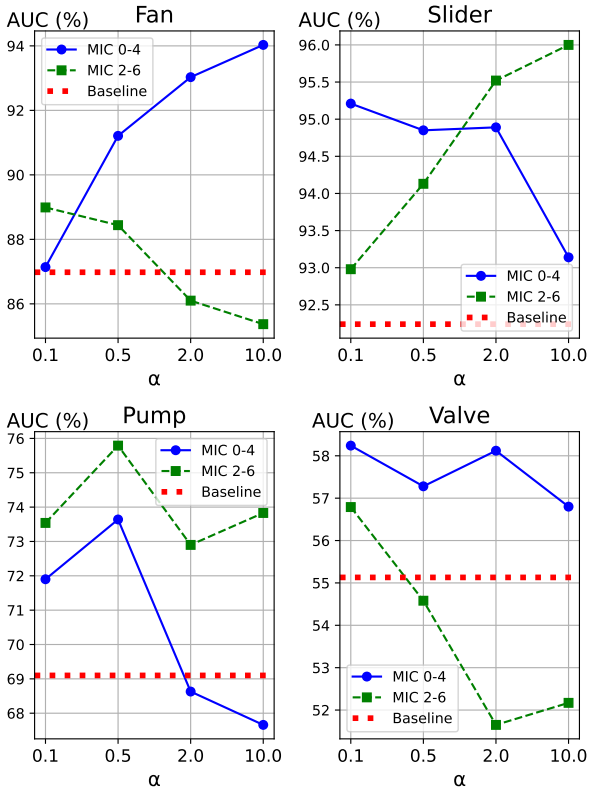

Figure 5. AUC values for each machine according to the microphone pair and the α value.

tunities to experiment with additional channels and microphone placements. These advancements pave the way for more robust and adaptable approaches to anomalous sound detection in industrial settings.

5. ACKNOWLEDGMENTS

This work has been supported by Grant TED2021-131003B-C21 funded by MCIN/AEI/10.13039/501100011033 and by the "EU Union NextGenerationEU/PRTR", as well as by Grant PID2022-137048OB-C41 funded by MICIU/AEI/10.13039/501100011033 and "ERDF A way of making Europe". Finally, the authors acknowledge as well the Artemisa computer resources funded by the EU ERDF and Comunitat Valenciana, and the technical support of IFIC (CSIC-UV).

6. REFERENCES

[1] J. A. Dar, K. K. Srivastava, and A. Mishra, "Lung anomaly detection from respiratory sound database (sound signals)," *Computers in Biology and Medicine*, vol. 164, p. 107311, 2023.

[2] 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.

[3] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for

machine condition monitoring: A domain generalization baseline," in *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 191–195, IEEE.

[4] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques."

[5] S. Ahmad, K. Styp-Rekowski, S. Nedelkoski, and O. Kao, "Autoencoder-based condition monitoring and anomaly detection method for rotating machines," in *2020 IEEE International Conference on Big Data (Big Data)*, pp. 4093–4102, 2020.

[6] F. Lachekhab, M. Benzaoui, S. A. Tadjer, A. Bensmaine, and H. Hamma, "Lstm-autoencoder deep learning model for anomaly detection in electric motor," *Energies*, vol. 17, no. 10, 2024.

[7] M. D. Chinnasamy, M. Sumbwanyambe, and T. S. Hlalele, "Acoustic anomaly detection of machinery using autoencoder based deep learning," in *2024 32nd Southern African Universities Power Engineering Conference (SAUPEC)*, pp. 1–6, 2024.

[8] S. Yan, H. Shao, Y. Xiao, B. Liu, and J. Wan, "Hybrid robust convolutional autoencoder for unsupervised anomaly detection of machine tools under noises," *Robotics and Computer-Integrated Manufacturing*, vol. 79, p. 102441, 2023.

[9] 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 *2020 IEEE Eighth International Conference on Communications and Electronics (ICCE)*, pp. 313–318, 2021.

[10] A. Jiang, W.-Q. Zhang, Y. Deng, P. Fan, and J. Liu, "Unsupervised anomaly detection and localization of machine audio: A gan-based approach," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[11] S. Y. Lee, J. Chang, and S. Lee, "Deep learning-based method for multiple sound source localization with high resolution and accuracy," *Mechanical Systems and Signal Processing*, vol. 161, p. 107959, 2021.

[12] V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, "Binaural speech enhancement using deep complex convolutional transformer networks," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 681–685, 2024.

[13] K. Imoto and N. Ono, "Spatial cepstrum as a spatial feature using a distributed microphone array for acoustic scene analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1335–1343, 2017.

[14] 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."

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.