Correlation-Based Filtering for Unsupervised Anomalous Sound Detection

Andrin Bürli, Sami Hamdan, Iason Kastanis

Centre Suisse d'Électronique et Microtechnique
Predictive Analytics Group, Switzerland
andrin.buerli@csem.ch, sami.hamdan@csem.ch

***Abstract*—Unsupervised anomalous sound detection (ASD) under domain shift remains a key challenge for real-world deployment. We introduce a two-stage "first-shot" pipeline for DCASE 2025 Task 2 that leverages optional clean-only or noise-only supplemental recordings to improve robustness to unseen background noises. First, a correlation-based filter is trained separately on clean or noise data, separating each test mixture $x = C + N + A$ into a cleaner signal $x' = C + A$. Second, a mel-spectrogram autoencoder, augmented with SMOTE and mixup on x', detects anomalies. On the development set, our method achieves a high SI-SDR for the separation task and improves the detection metrics for three out of seven components compared to the baseline. These results validate that assuming statistical independence between machine sound, background noise, and anomalies can enhance first-shot ASD. Future work will explore automated correlation estimation and integration with more advanced anomaly detection methods for the second stage.**

***Index Terms*—anomalous sound detection, signal correlation, DCASE, source separation, audio**

1. INTRODUCTION

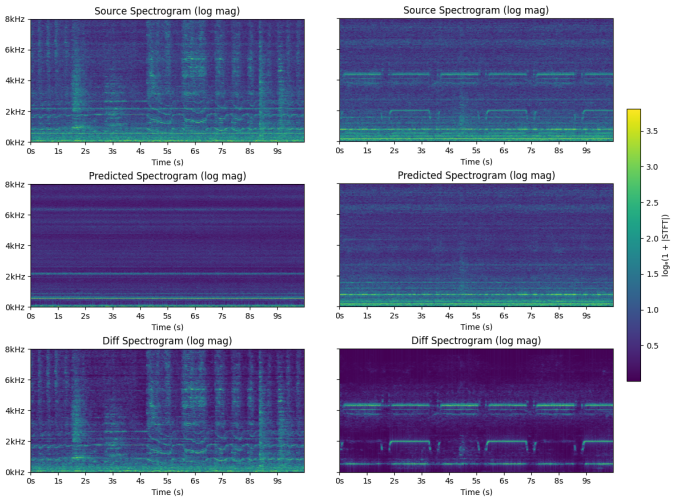
Anomalous sound detection (ASD) has emerged as a critical technology for non-intrusive monitoring of industrial machinery, enabling early warning of mechanical faults through audio analysis [1], [2]. Unsupervised ASD, which relies solely on normal-condition recordings, was first standardized in the DCASE 2020 Challenge Task 2 to address the scarcity and diversity of anomalous examples in real factories [3]. Subsequent editions have progressively incorporated domain-shift and "first-shot" scenarios, in which systems must generalize to unseen operating conditions or entirely new machine types without task-specific tuning [3], [4].

Building on the first-shot unsupervised ASD framework of DCASE 2023 and 2024, the 2025 Task 2 challenge retains the requirement to train exclusively on normal data and to detect anomalies under unknown domain shifts, while introducing optional use of clean-only or noise-only supplementary recordings [4]. Participants must also handle completely novel machine types at evaluation, with no access to anomalous test data for hyperparameter tuning. This "first-shot" setting reflects real-world constraints where rapid deployment precludes exhaustive data collection or manual calibration.

We propose a two-stage pipeline for first-shot ASD: (1) a correlation-based separator that, given clean-only or noise-only supplemental data, filters each test mixture $x = C + N + A$ into $x' = C + A$ as depicted in Figure 1; (2) a mel-spectrogram autoencoder, augmented with SMOTE and mixup trained on x', to detect anomalies. By leveraging correlation-based filtering, our method enhances robustness to unseen background noises in the DCASE 2025 Task 2 setting.

2. METHOD

We denote clean machine sound by C, background noise by N, and anomalous sound by A. Artificial noise augmentations, N_A, are sampled from diverse sources. We use $\rho(S_1, S_2)$ as the correlation between two signals S_1 and S_2, where the threshold ε denotes a significant correlation between them. In our two-stage methodology, a first step trains a filtering model which can separate the mixture $x = C + N + A$ into $x' = C + A$. The second step then involves training an anomaly detection model based on the filtered x'.



(a) Bearing (b) Slider

Fig. 1: Correlation-based filtering for two components. In (a) we have supplemental $= C$, thus we perform machine sound extraction. In (b) we have supplemental $= N$, thus we perform noise extraction.

2.1. Correlation-based Filtering

In the DCASE challenge 2025, we are provided with additional supplemental data which consists of either C or N. For both cases we developed separate filtering strategies. For the evaluation of the filtering quality, we report the scale-invariant signal-to-distortion ratio (SI-SDR), which is commonly used in source separation tasks [5].

2.1.1. Machine Sound Extraction: If we are provided with supplemental data containing the clean machine sound C, we train a source-separation network

$$f_\theta(C + N_A) \approx C \tag{1}$$

to recover the clean sound C from noise-augmented inputs $C + N_A$. At inference time on mixture x this will allow us to filter:

$$\widehat{y} = f_\theta(x) \approx C + A = x' \tag{2}$$

For this filtering to work we introduce the following assumptions:

1.1 $\rho(C, N_A) < \varepsilon$ (artificial noise uncorrelated with C)
1.2 $\rho(C, N) < \varepsilon$ (background noise uncorrelated with C)
1.3 $\rho(C, A) > \varepsilon$ (anomalies strongly correlated with C)
1.4 $\rho(N, A) < \varepsilon$ (anomalies uncorrelated with N)
1.5 $\rho(C_{\text{source}}, C_{\text{target}}) > \varepsilon$ (machine sound of source is strongly correlated with target)

2.1.2. Noise Extraction: If we are provided with supplemental data containing the background sound N, we train a source-separation network

$$f_\theta(N + N_A) \approx N \tag{3}$$

to extract background noise N from $N + N_A$. At inference time on mixture x this will allow us to filter:

$$\widehat{y} = f_\theta(x) \approx N \rightarrow x - \hat{y} = x' \tag{4}$$

For this filtering to work we introduce the following assumptions:

2.1 $\rho(N, N_A) < \varepsilon$ (artificial noise uncorrelated with N)
2.2 $\rho(A, N_A) < \varepsilon$ (artificial noise uncorrelated with A)
2.3 $\rho(C, N) < \varepsilon$ (background noise uncorrelated with C)
2.4 $\rho(N, A) < \varepsilon$ (anomalies uncorrelated with N)
2.5 $N_{\text{source}} = N_{\text{target}}$ (background sound of source is equal to target)

Assumptions 2.1, 2.2, and 2.5 are new; 2.3 and 2.4 overlap with 1.2 and 1.4, respectively.

2.2. Anomaly Detection

Once we have x' we can theoretically use any of the methods presented in the DCASE challenges 2020-2024, ranging from outlier exposure to inlier modeling and a large diversity of combinations of the two [6]–[9]. We choose to use a similar approach as the baseline of the 2025 challenge, consisting of an autoencoder based on Mel-spectrograms. Additionally, we employ SMOTE [10] for oversampling the target domain and mixup [11] to augment our data. For anomaly detection, we evaluate using the area under the ROC curve (AUC) and partial AUC (pAUC), following the official DCASE challenge metrics [4].

3. EXPERIMENTAL SETUP

We adhere to the DCASE 2025 Task 2 protocol [4]. The development dataset provides training and test splits for seven machines: Valve, Bearing, ToyCar, ToyTrain, Slider, Gearbox and Fan, where we have supplemental C in the first three and N in the others. For each component, we first train a correlation-based filter model with the strategy depending on the provided supplemental data. The model is a standard U-Net with input and output being the complex spectrograms of the respective signals using a 64-ms window and 32-ms hop size [12]. U-Net has proven effective for source separation, especially when the thresholds ε in assumptions 1.1 and 2.1 are small. We use a batch size of 32 and a learning rate of 0.0005 to optimize over a multi-resolution STFT loss [13] for 300 epochs with early stopping. To counteract potential violations of assumptions 1.1 and 2.1, we sweep over various SNR ranges and N_A sources and choose the run resulting in the highest adjusted SI-SDR ($=$ SI-SDR $- \mathbb{E}[\text{SNR}]$) on a 10% holdout validation set, to ensure we cover realistic SNR ratios that are not known in advance.

- N_A sources: {AudioSet full, AudioSet no tools, AudioSet only tools, DCASE Clean (supplemental clean-only), DCASE Noise (supplemental noise-only)},
- SNR windows: [-30,30], [-10,30], [-10,10], [-5,5] dB.

The anomaly detection autoencoder is trained with very similar parameters as the DCASE 2025 baseline. We use a 64-ms window with a 128-bin mel spectrogram over five consecutive windows as a feature vector. The encoder-decoder architecture is a symmetric MLP. We train the model over 100 epochs with a learning rate of 0.001 and a batch size of 64.

4. RESULTS

We evaluate our two-stage pipeline on the DCASE 2025 Task 2 development set in three parts: correlation-based filtering on the development data, filtering performance on the additional evaluation data, and anomaly detection on the development set.

Component	SNR [dB]	N_A Source	SI-SDR [dB]
Valve	[-10, 10]	AudioSet full	11.4
ToyTrain	[-10, 10]	DCASE Clean	6.3
ToyCar	[-5, 5]	AudioSet full	5.8
Slider	[-5, 5]	DCASE Clean	6.9
Gearbox	[-5, 5]	DCASE Clean	5.7
Fan	[-5, 5]	DCASE Clean	7.4
Bearing	[-5, 5]	AudioSet full	9.1

Table 1: Best Adjusted SI-SDR results for correlation-based filtering per component in development dataset

First, Table 1 reports the optimal filtering settings and adjusted SI-SDR for each of the seven development components. Valve achieves the highest SI-SDR of 11.4 dB using full AudioSet noise at ±10 dB, while Bearing achieves 9.1 dB under a narrower ±5 dB range with the same noise source. ToyTrain (6.3 dB) and ToyCar (5.8 dB) similarly leverage wider SNR windows (±10 dB and ±5 dB) with DCASE Clean or AudioSet full augmentations, reflecting their varied spectral content. The remaining components Slider (6.9 dB), Gearbox (5.7 dB), and Fan (7.4 dB) attain the best separation under narrow (±5 dB) clean-only noise, indicating limited noise variability suffices for these cases.

Component	SNR [dB]	N_A Source	SI-SDR [dB]
AutoTrash	[-5, 5]	AudioSet full	15.29
BandSealer	[-10, 10]	DCASE Clean	6.79
CoffeeGrinder	[-5, 5]	DCASE Clean	11.38
HomeCamera	[-5, 5]	DCASE Clean	12.07
Polisher	[-10, 10]	AudioSet full	5.82
ScrewFeeder	[-5, 5]	AudioSet no tools	8.84
ToyPet	[-5, 5]	DCASE Clean	9.16
ToyRCCar	[-5, 5]	DCASE Clean	8.67

Table 2: Adjusted SI-SDR for correlation-based filtering per component in additional training dataset

Next, Table 2 presents SI-SDR results on eight novel components in the additional evaluation set. Here, SI-SDR ranges from 5.82 dB (Polisher) up to 15.29 dB (AutoTrash), with most components favoring ±5 dB clean or full-AudioSet noise. This consistency confirms that our correlation-based filter generalizes effectively to unseen machine types in a first-shot scenario.

Component	Baseline	Best 2024	Unfiltered	Filtered
Valve	0.611	0.771	0.669	**0.848**
ToyTrain	0.557	**0.651**	0.590	0.564
ToyCar	0.567	**0.594**	0.588	0.405
Slider	0.561	0.593	0.542	**0.600**
Gearbox	0.553	**0.704**	0.547	0.566
Fan	0.499	**0.639**	0.541	0.545
Bearing	0.598	0.691	0.582	**0.734**
hmean	0.5617	**0.6582**	0.5771	0.5775

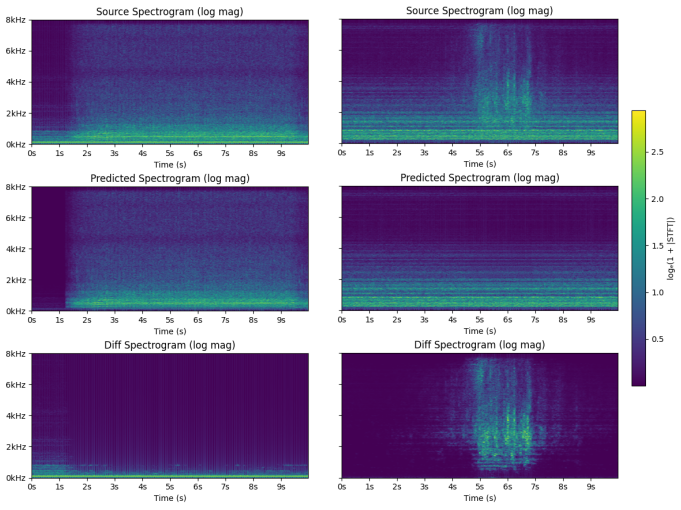
Table 3: Development dataset detection results. Scores correspond to the harmonic mean of AUC and pAUC. Best 2024 corresponds to [14], best results per row are highlighted in bold.

Finally, Table 3 compares anomaly detection metrics before (”Unfiltered”) and after filtering (”Filtered”), alongside the baseline and the Best 2024 system. We would expect ”Filtered” to be similar or better than ”Unfiltered” if the filtering indeed transforms our mixture $x = C + N + A$ to $f_\theta(x) = C + A = x'$. After filtering, Valve improves from 0.669 to 0.848 (+0.179), Bearing from 0.582 to 0.734 (+0.152), Slider from 0.542 to 0.600 (+0.058) and Gearbox from 0.547 to 0.566 (+0.019), demonstrating that isolating $C + A$ enhances anomaly detection as expected.

Conversely, ToyTrain (0.590 $\rightarrow$ 0.564, –0.026) and ToyCar (0.588 $\rightarrow$ 0.396, –0.183) degrade, indicating their anomalies may not align with our independence assumptions. Overall, the harmonic mean across components increases slightly from 0.5771 to 0.5775 (+0.0004), evidencing modest benefits of filtering on average across all machines, which is mostly due to the bad results on the ToyCar and ToyTrain components.

5. DISCUSSION

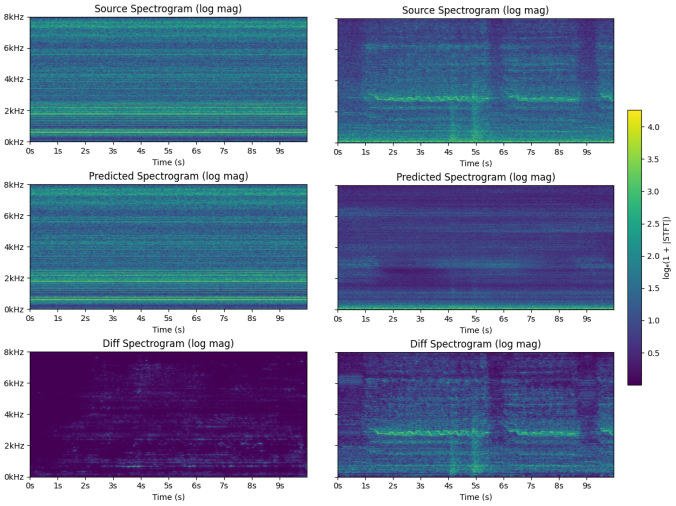
The experimental results demonstrate that our correlation-based filtering effectively enhances anomaly detection when the underlying independence assumptions hold. For components such as Valve, Bearing, and Slider, the filter succeeded in isolating the machine signal plus anomaly, leading to clear gains in anomaly detection (Table 3, Figure 1). This indicates that, for these machines, (1) background noise and artificial augmentations remain uncorrelated with the clean sound ($\rho(C, N_A) < \varepsilon$ and $\rho(N, N_A) < \varepsilon$), and (2) anomalous events retain sufficient correlation with the machine signature ($\rho(C, A) > \varepsilon$) to survive filtering.



(a) ToyCar (b) ToyTrain

Fig. 2: Both components in (a) and (b) produce non-stationary sounds that are correctly filtered. However, they may violate assumptions 1.3, 1.4, and 2.4 because the anomalous sound may be weakly correlated with the machine sound, for example, by only occurring during the ramp-up or ramp-down phase.

In contrast, ToyTrain and ToyCar exhibit performance degradations after filtering, with detection scores falling by 0.026 and 0.183 respectively. Their non-stationary operating cycles with ramp-up, steady, and ramp-down phases appear to violate the assumption that anomalies strongly co-vary with the baseline machine sound. As a result, the filter may remove or attenuate anomalous components along with noise, harming detection (Figure 2). Introducing additional transformations such as windowing could help with this issue but requires further work. Fan and Gearbox exhibit only modest detection gains after filtering, indicating partial alignment with our independence assumptions. We attribute this to an under-representation of background noise in the supplemental data (see Figure 3): for Fan, the supplemental recordings contain only stationary noise which might be easier to detect in STFT, whereas the development and evaluation sets also include non-stationary events such as hammering and grinding. Consequently, the filtering model cannot learn to suppress these dynamic noise components, leaving residual interference in x' and limiting the achievable improvement. We made very similar observations for Gearbox.



(a) Sample of fan supplemental background noise (b) Sample of fan training data

Fig. 3: In (a) we see a representative sample from supplemental data for Fan, which is stationary. In (b) we can see that the actual training data contains obvious non-stationary background events, such as grinding. The correlation-based filter model does not remove these events because it has never encountered them before.

On the additional evaluation set, the filter generalizes effectively to eight novel components, yielding SI-SDR scores between 5.82 dB and 15.29 dB (Table 2). Although absolute separation quality varies with machine-noise spectral overlap, the consistent performance across unseen machines confirms the robustness of our first-shot filtering approach.

6. CONCLUSION

We have presented a two-stage “first-shot” pipeline for unsupervised anomalous sound detection, combining correlation-based filtering with a mel-spectrogram autoencoder. By grid-searching SNR windows and noise-augmentation sources, our method adaptively separates each mixture into machine-plus-anomaly signals before a simple reconstruction-based error detection. On the DCASE 2025 Task 2 development set, filtering improved detection metrics for the majority of components. On the eight unseen machines we find a similar separation performance range as for the development dataset using the same hyperparameter grid. Future work will explore automated estimation of signal correlations to select augmentations per machine and integration with more sophisticated anomaly detectors that can tolerate partial assumption violations.

REFERENCES

[1] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "Non-linear prediction with lstm neural networks for acoustic novelty detection," *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2015.

[2] T. Pereira and N. Nunes, "Anomaly detection in industrial shop-floor machines using audio and vibration signals," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 854–858.

[3] Y. Nishida, V. Saponaro, M. Dorfer, and N. Ono, "First-shot unsupervised anomalous sound detection challenge: Overview and baseline system," in *Proceedings of the DCASE 2024 Workshop*, 2024, pp. 150–155.

[4] DCASE 2025 Challenge Task 2 organizers, "First-shot unsupervised anomalous sound detection for machine condition monitoring," Online, 2025.

[5] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[6] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," 2020. [Online]. Available: https://arxiv.org/abs/2006.05822

[7] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on dcase 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," 2021. [Online]. Available: https://arxiv.org/abs/2106.04492

[8] 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," 2022. [Online]. Available: https://arxiv.org/abs/2206.05876

[9] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on dcase 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," 2023. [Online]. Available: https://arxiv.org/abs/2305.07828

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[11] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[12] A. Jansson, R. M. Bittner, N. Montecchio, and T. Weyde, "Learned complex masks for multi-instrument source separation," 2021. [Online]. Available: https://arxiv.org/abs/2103.12864

[13] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[14] A. Jiang, Q. Hou, J. Liu, P. Fan, J. Ma, C. Lu, Y. Zhai, Y. Deng, and W.-Q. Zhang, "Thuee system for first-shot unsupervised anomalous sound detection for machine condition monitoring," *Proceedings of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events, Tampere, Finland*, pp. 20–22, 2023.