SVD DECOMPOSITION WITH AUTOENCODERS FOR DCASE 2025 TASK 2

Technical Report

Vladimir Igoshin, Vsevolod Kleshchenko, Dmitry Chirkov, Mark Mirolyubov, Mihail Petrov, Igor Lobanov

ITMO University
School of Physics and Engineering,
ITMO University, Saint Petersburg 197101, Russia

ABSTRACT

In this work, we address the problem of single-channel sound anomaly detection by leveraging Singular Value Decomposition (SVD) as a feature extraction and dimensionality reduction technique. Specifically, we apply SVD across the entire dataset of spectrograms and retain only a limited number of dominant components to represent the input signals in a compact latent space. We evaluate two autoencoder-based models on the reduced representations. First one is a challenge baseline autoencoder trained on the low-dimensional features obtained from SVD. Second is transformer-inspired autoencoder that integrates a convolution layer and an attention mechanism to better capture temporal structures indicative of anomalous behavior.

***Index Terms*—** SVD, AE, Transformer

1. INTRODUCTION

Recently, monitoring the condition of various types of equipment based on their emitted acoustic signals has emerged as a highly promising area of research. In this report, we present the results of our adapted algorithms developed to address Task 2 "First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring" of the DCASE 2025 challenge [1]. The task involves classifying 10-second audio recordings of different machine types as either normal or anomalous. A key challenge lies in the fact that the algorithms are trained solely on recordings of normal machine sounds, while test data may originate from different domains and contain arbitrary background noise. A detailed description of the task [1] and the datasets [2, 3, 4] is available on the official challenge webpage.

In this report, we provide a brief overview of the algorithms we employed and the results obtained for the proposed task. Specifically, we utilized two adapted methods based on an autoencoder and a transformer architectures, both operating on preprocessed data obtained via principal component extraction from input spectrograms.

2. METHODS

2.1. Features extraction

In the initial phase, waveforms were transformed to spectrograms using 1024 samples Hann window and 512 samples hop using PyTorch implementation [5]. The phase information was discarded, only amplitudes were analysed. All further computations were performed on natural logarithms of amplitudes. Stationary noise was partially removed from the spectrograms by subtracting noise power spectrum, computed for each sound clip separately. The noise spectrum was estimated by taking weighted average over time axis, where the weight is inversely proportional to the signal volume in the corresponding window.

Main features are extracted using Principal Component Analysis (PCA) implemented via a truncated singular-value decomposition (SVD), also known as Karhunen-Loève Transform or Empirical Orthogonal Function (EOF) analysis. If machine only source is provided in supplementary data, only machine sound is used for features selection, otherwise the training data is used. The corresponding spectrograms are concatenated to single matrix A along time axis, then reduced SVD for the matrix is computed $A = UDV$. The orthogonal matrix U defines mapping from newly selected features to the frequency domain. Main features were selected to correspond to columns of U matching 32 largest singular values; denote the obtained 512×32 matrix U'. Then spectrogram of every sound clip is transformed to a feature-gram by multiplying the spectrogram by U'.

2.2. Models

2.2.1. Baseline autoencoder (SVD AE)

As a first step, we tested the baseline model [6] using our proposed data representation to validate the suitability. Following the baseline approach, five consecutive time frames of length 32 were extracted from each transformed spectrogram. Based on this representation, we implemented an autoencoder architecture consisting of an encoder with layer dimensions of 160–128–128–128–8. Each linear layer was followed by a batch normalization layer and a ReLU activation function. The decoder was structured symmetrically.

The total number of trainable parameters in the model is 144,568. During training, we used batches of size 256 and optimized the model using the Adam optimizer with a learning rate of 0.001 and mean squared error (MSE) as the loss function. More details on description and the original implementation can be found in baseline model description [6] and related sources [1, 7].

The model was trained separately for each machine type, with the objective of minimizing the reconstruction error on normal data. For the evaluation dataset [4], the anomaly detection threshold was selected based on the distribution of reconstruction errors in the normal data in additional training dataset [3]. For the development dataset [2], the threshold on the reconstruction error was determined using the receiver operating characteristic curve (ROC) by maxi-

Thanks to NIRSII for funding.

mizing the difference between the true positive rate and the false positive rate.

2.2.2. *Transformer-inspired autoencoder (SVD TransE)*

The second model is designed to leverage temporal context for more accurate frame prediction within a sequence of feature embeddings. The primary objective of this model is to reconstruct the central frame of a temporal window based on its surrounding context. Specifically, the input to the model consists of a fixed-length sequence of 11 consecutive frames, where the central frame is intentionally excluded from the input and used only as the prediction target. The surrounding frames are then passed through a 1D convolutional layer, which captures short-term temporal dependencies. This convolution is applied across the temporal axis with kernel size of 5 and 128 output channels. A GELU activation function is applied to the convolutional outputs, followed by batch normalization.

Following the convolutional layer, the transformed frame representations are passed through a linear layer to reduce the dimentiality to 16. These embeddings are then processed by a Transformer encoder composed of 2 layers of multi-head self-attention and feed-forward networks with a hidden dimension of 24. This component is responsible for modeling the global dependencies across the temporal sequence, allowing the model to learn rich contextual relationships between the remaining frames in the input.

To reconstruct the missing central frame, the model employs a learnable query vector for the masked central frame. This query is used in a multi-head attention mechanism, while the output of the Transformer encoder serves as both the key and value. Finally, the resulting vector is passed through a linear decoder layer to upscale it back into the original embedding space of size 32. The output is compared to the ground-truth central frame using a mean squared error loss.

3. RESULTS

The performance of the proposed algorithms for the DCASE 2025 Task 2 on the development dataset is presented in Table 1. For each machine type in the test dataset, we report the values of the area under the ROC-curves (AUC) for both the source and target domains and compare them with the baseline model. Both of our proposed methods SVD AE and SVD TransE show a clear improvement over the baseline for the ToyTrain in the source domain. Additionally, the SVD TransE method achieves a strong result for the valve in the source domain, although it is inferior to the baseline model in the target domain.

Table 1: Anomaly detection results (%) for different machine types

	Metric	**Baseline MSE**	**SVD AE**	**SVD TransE**
ToyCar	AUC(source)	71.05	40.84	39.60
	AUC(target)	53.52	48.32	45.48
ToyTrain	AUC(source)	61.76	76.96	70.24
	AUC(target)	56.46	55.32	56.80
bearing	AUC(source)	66.53	50.20	55.20
	AUC(target)	53.15	48.80	50.12
fan	AUC(source)	70.96	58.80	54.12
	AUC(target)	38.75	51.84	56.40
gearbox	AUC(source)	64.80	55.84	52.56
	AUC(target)	50.49	55.16	54.84
slider	AUC(source)	70.10	48.52	45.48
	AUC(target)	48.77	47.32	42.20
valve	AUC(source)	63.53	53.52	80.80
	AUC(target)	67.18	47.84	50.80

4. REFERENCES

[1] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and Discussion on DCASE 2025 Challenge Task 2: First-shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," *arXiv*, June 2025.

[2] ——, "Dcase 2025 challenge task 2 development dataset," Apr. 2025. [Online]. Available: https://doi.org/10.5281/zenodo.15097779

[3] ——, "Dcase 2025 challenge task 2 additional training dataset," May 2025. [Online]. Available: https://doi.org/10.5281/zenodo.15392814

[4] ——, "Dcase 2025 challenge task 2 evaluation dataset," June 2025. [Online]. Available: https://doi.org/10.5281/zenodo.15519362

[5] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *arXiv*, Dec. 2019.

[6] 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)*. IEEE, 2023, pp. 191–195.

[7] "Dcase 2023 task 2 baseline autoencoder," https://github.com/nttcslab/dcase2023_task2_baseline_ae, 2023.