ANOMALOUS SOUND DETECTION IN INDUSTRIAL MACHINERY USING NORMALIZING FLOW-BASED DEEP LEARNING

Technical Report

Natalia P. García-de-la-Puente[1*], *Fran Pastor-Naranjo*[1], *Miguel López-Pérez*[1], *Gema Piñero*[2], *Valery Naranjo*[1]

[1] HumanTech, Universitat Politècnica de València, Spain, napegar@upv.es
[2] ITEAM, Universitat Politècnica de València, Spain, gpinyero@iteam.upv.es

ABSTRACT

This technical report describes our approach for Task 2 of the DCASE 2024 Challenge. This task aims to develop an anomalous sound detection (ASD) system to determine whether the sound emitted from a target machine is normal or anomalous. To tackle this task, we propose an unsupervised deep learning model based on a normalizing flow architecture. Our framework consists of a pre-trained encoder (WideResNet50) and a multi-scale generative decoder to estimate the log-likelihoods of feature vectors. The input to the model is an image comprising four different time-frequency representations of the sounds (Mel spectrogram, CQT-chroma, tonnetz, and spectral contrast) together with five 1D characteristics computed along the time index. All the 2D and 1D features are concatenated in the frequency dimension, resulting in an image 158 pixels high, with their width depending on the duration of the sounds.

***Index Terms*—** Anomalous sound detection, C-Flow, Mel spectrogram, industrial machinery.

1. INTRODUCTION

The DCASE 2024 Challenge Task 2 [1] focuses in Anomalous Sound Detection (ASD) for Machine Condition Monitoring. It is a first-shot unsupervised problem because it involves training a model using a limited number of machines from its machine type, and the source and target domain data are imbalanced. This year, only a few machines have attribute information. The system must perform effectively both with and without this information.

The task offers two baseline methods: a standard autoencoder and a method that uses Mahalanobis distance autoencoder. The first excels in ASD but struggles with domain generalization, and the second performs well in the source domain, but does not work properly in target domain.

Our approach employs conditional normalizing flows using memory-efficient architecture. The framework imposes multivariate Gaussian (MVG) prior to the parameters and the feature vectors extracted from the feature maps. After this, it measures the Mahalanobis distance of a particular feature vector to its MVG distribution. Finally, it uses generative probabilistic models (flows) based on layered transformations to estimate the exact likelihood of any arbitrary distribution.

2. DCASE2024 TASK 2 DATASETS

The organizers provided three datasets [2, 3] of machine sounds, and all of them are mono recordings. The first dataset is the Development dataset comprising 1000 normal and anomalous operating sounds from seven types of machines. Their sounds are 10s long except for two of the machines that are 12s long, and the seven classes include source and target sounds. Secondly, the Additional training dataset consists of 1000 sounds from nine machines, all of them different from the machines of the Development dataset. Their duration varies from 6 to 10s long and they also include source and target labels, but all of them are classified as normal. The Evaluation dataset has 200 recordings from the same machines of the Additional training dataset, but it comprises unlabeled normal and anomalous sounds.

2.1. Audio features extraction

We have extracted several audio features to form the input matrix of the model described in section 3 inspired by the ensemble of audio features used in [4, 5] to classify environmental sounds. For the time processing we have used a hann window of 64ms (sampling frequency $f_s = 16$kHz). The extracted features at each time frame have been:

- The log-energy of the Mel spectrogram of 128 bins.

*This work has been funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe" through Grants PID2021-124280OB-C21, PID2022-140189OB-C21, JDC2022-048318-I and by EU Horizon Europe through Grant Agreement No 101057404.

- The constant-Q transform (CQT) chroma [6] resulting in 12 values.
- The spectral contrast [7] of 7 octave bands.
- The the tonal centroid features (tonnetz) [8] resulting in 6 values.
- The root mean square value.
- The spectral centroid.
- The spectral bandwidth.
- The spectral flatness.
- The zero-crossing rate.

All the audio features were concatenated along the y-axis forming an input matrix of dimensions $[158 \times N_f]$, being N_f the number of time frames, which depends on the particular duration of each type of machine. All the audio features have been computed with librosa [9].

As a final remark, the CQT-chroma, tonnetz and spectral contrast features are often used when processing music audio signals since they can represent and discriminate harmonic content. However, some types of machinery noises are harmonic and we have considered worthy to include them in this work.

3. MODEL DESCRIPTION

Normalizing flows are unsupervised generative models. They can serve as a suitable estimator of probability densities to detect anomalies because they are able to learn transformations between data distributions. Compared to other methods, they are promising because of their ability to generalize and high inference efficiency [11].

We utilize CFLOW [10], which consists of an encoder for feature extraction and a decoder for likelihood estimation. As shown in Figure 1, our experiments use the same multiscale feature pyramid pooling to capture details at different resolutions within the audio features concatenated. The encoder employed is ImageNet-pretrained WideResnet-50.

After this stage, the aim is to fit different densities with conditional normalizing flow framework and independent decoder models. The reason why multiple decoders are used is because of the multi-scale pyramid pooling setup feature. The decoder incorporates spatial prior, generating a conditional vector that contains *sin* and *cos* harmonics to encode the position, and a sequence of coupling layers. Each coupling layer is fully connected, with softplus activation and permutations of the output vector.

Finally, both the encoder and the decoders have translational-equivariant convolutional architectures, because they use kernel parameter sharing. After estimating the probability density functions and transforming them into Gaussian distributions, the classification decision is based on an adaptive theshold.

	Method	Baseline	CFlow
ToyCar	AUC(source)	66.98%	$\mathbf{69.13\% \pm 4.87\%}$
	AUC(target)	33.75%	$\mathbf{53.25\% \pm 6.61\%}$
ToyTrain	AUC(source)	**76.63%**	$60.56\% \pm 7.31\%$
	AUC(target)	46.92%	$\mathbf{61.3\% \pm 11.24\%}$
Bearing	AUC(source)	62.01%	$\mathbf{74.06\% \pm 3.19\%}$
	AUC(target)	61.4%	$\mathbf{66.09\% \pm 11.73\%}$
Fan	AUC(source)	**67.71%**	$67.45\% \pm 1.37\%$
	AUC(target)	**55.24%**	$45.52\% \pm 15.16\%$
Gearbox	AUC(source)	**70.4%**	$64.96\% \pm 2.21\%$
	AUC(target)	**69.34%**	$62.52\% \pm 4.14$
Slider	AUC(source)	66.51%	$\mathbf{72.42\% \pm 4.28\%}$
	AUC(target)	**56.01%**	$51.96\% \pm 6.82$
Valve	AUC(source)	51.07%	$\mathbf{72.05\% \pm 6.12\%}$
	AUC(target)	46.25%	$\mathbf{56.47\% \pm 6.32\%}$

Table 1: Average AUC values through three independent runs. Development dataset machines for the baseline MSE and the proposed method.

4. RESULTS AND DISCUSSIONS

Table 1 presents the results of our model compared to the baseline described in [12]. Our audio features with CFlow model shows better performance on the target domain. It is important to highlight the training (Avg. 9 minutes) and inference times (Avg. 5.5 seconds), since CFlow converges to similar results but with less computation when using MVG assumptions. It is interesting to note that using the proposed method is not a benefit for the fan and gearbox machines, so that a line of research should be opened to include data such as attributes that would allow the model to better understand the normality of these samples.

5. 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 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2406.07250*, 2024.

[2] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1–5.

[3] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII

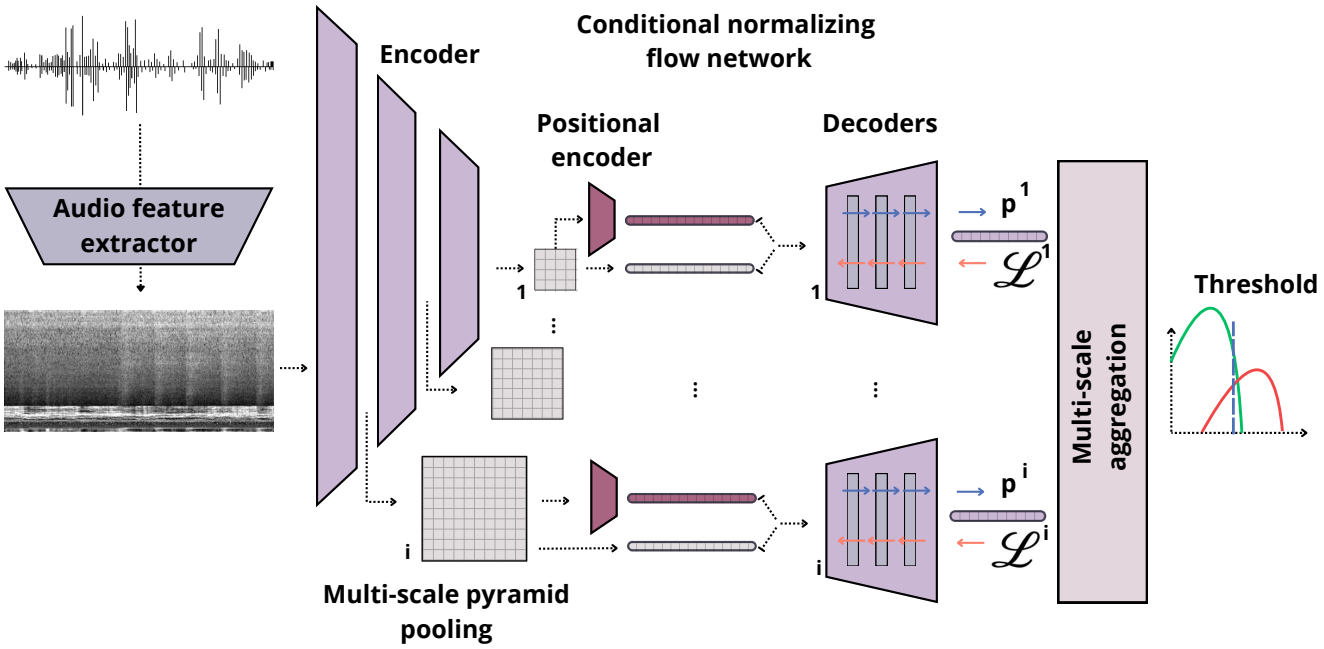

Figure 1: Framework proposed. Audio feature extractor with CFLOW [10] adaptation.

DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[4] Z. Xing, E. Baik, Y. Jiao, N. Kulkarni, C. Li, G. Muralidhar, M. Parandehgheibi, E. Reed, A. Singhal, F. Xiao, and C. Pouliot, "Modeling of the latent embedding of music using deep neural network," 2017.

[5] Y. Su, K. Zhang, J. Wang, D. Zhou, and K. Madani, "Performance analysis of multiple aggregated acoustic features for environment sound classification," *Applied Acoustics*, vol. 158, p. 107050, 1 2020.

[6] M. Bartsch and G. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[7] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," in *Proceedings. IEEE International Conference on Multimedia and Expo*, 2002, pp. 113–116 vol.1.

[8] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*, 2006, p. 21–26.

[9] B. McFee, M. McVicar, D. Faronbi, I. Roman, M. Gover, and S. B. et al., "librosa/librosa: 0.10.2.post1," May 2024. [Online]. Available: https://doi.org/10.5281/zenodo.11192913

[10] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 98–107.

[11] Y. Cui, Z. Liu, and S. Lian, "A survey on unsupervised visual industrial anomaly detection algorithms," *arXiv e-prints*, pp. arXiv–2204, 2022.

[12] 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)*, 2023, pp. 191–195.