# ANOMALY MACHINE DETECTION ALGORITHM BASED ON SEMI VARIATIONAL AUTO-ENCODER OF MEL SPECTROGRAM

Technical Report

Ke Tian[1], Guoheng Fu[1], Shengchen Li[1], Gang Tang[2], Xi Shao[3]

[1] Beijing University of Posts and Telecommunications
tianke3366@gmail.com, {fgh, shengchen.li}@bupt.edu.cn
[2] Beijing University of Chemical Technology
tanggang@mail.buct.edu.cn
[3] Nanjing University of Posts and Telecommunications
shaoxi@njupt.edu.cn

## ABSTRACT

This report proposes a solution for Task 2 of IEEE DCASE data challenge 2020, which attempts to detect anomaly machines according to acoustic data. The proposed solution uses a semi variational auto-encoder. The term “semi” indicates that the resulting variational auto-encoder may not successfully reconstruct the input as the key task of the task is to distinguish the outlier samples according to a specific feature rather than reconstruct the input precisely. As a result, there are a few minor changes introduced by the provided baseline system, which set up a different training stop criteria and a different anomaly scoring system. By the proposed method, the use of different stop training criteria for an variational auto-encoder may help different objectives.

Index Terms— Variational Auto-encoder, Anomaly Detection, Acoustic Signal Processing

## 1. INTRODUCTION

This report proposes a submission of the task, “Unsupervised Detection of Anomalous Sounds for Machine Condition Monitoring”, in the 2020 edition of IEEE AASP Challenge on Detection and Classification of Acoustic Scene and Environment (DCASE). The task requires participants to identify anomaly of six types of machines according to the recording of machine operation audio with only sound of normal machines being used for training. The proposed system make minor changes to the provided baseline system with potential improvements of performance.

The baseline system uses an auto-encoder to establish a mapping between the mel spectrogram of the audio and an non-semantic representation in latent space, which is then reconstructed by a decoder that maps the resulting representation in latent space to a reconstructed mel spectrogram. The anomaly score of a piece of audio is then calculated by the loss between the original mel spectrogram and the reconstructed mel spectrogram.

Based on the baseline system, a few observations are made including the reconstruction loss changes at each epochs, the average distance of latent representation vectors between normal audio and abnormal audio, the entropy of dataset for the latent representations of normal audio and the likelihood of normal and abnormal latent vectors for the model of latent vector distributions. For an easier observation of the proposed features, the proposed system uses a variational auto-encoder instead of a simple auto-encoder, which fits the latent vectors to a Gaussian distribution.

With the observations of proposed features, the best performed auto-encoder that converges and has the lowest reconstruction loss is not necessary to produce the most distinguished features between normal and abnormal machines. As a result, the submitted system proposes a different stopping criteria of auto-encoder training compared with the baseline system to pursue a better performance. As a result, the proposed solution is named as a semi variational auto-encoder since the training process is stopped earlier before the loss criterion in the training process converges.

The remaining parts of the report is organised in the following way. The proposed system is firstly introduced together followed by the introduction of training stop criterion. The final results of development dataset are shown before a quick conclusion of this report.

## 2. PROPOSED SYSTEM

The proposed system uses mel spectrogram as the input of the system. A semi variational auto-encoder is then trained, where the training process is terminated by features other than reconstruction loss, much earlier before the engaged loss criterion in the training process.

### 2.1. Mel Spectrogram

The feature used in the system is mel spectrogram where 128 frequency bins are used as following the baseline. The window length is 1024 points with a hope size of 512 points, which also follow the baseline set-up.

### 2.2. Deep Neural Network

The proposed system uses a deep neural network, whose inputs and outputs are expected to be identical which are expected to be the mel spectrogram of five successive frames.

For easier analysis of latent space and more sensible distributions of latent vectors, Kingma and Welling [1] proposed an algorithm that fits latent vectors to a Gaussian-like model and sampling from the Gaussian-like models to reconstruct the inputs, which is named as variational auto-encoders.

The proposed system follows the exact architecture of variational auto-encoders. There are four layers of forward neural network (the `nn.Linear` layers in `PyTorch`) in the encoder, whose numbers of neurons are `128`, `128`, `64`, `64` respectively. The latent presentation of inputs are produced by a forward neural network containing `30` neurons thus the latent vectors have a dimension of 30.

With a Gaussian model with single Gaussian component, another set of latent vectors are then samples from the resulting model then passed into the decoder whose layers are the same of encoders with an inverse order.

The labels of output are set to the same of inputs, *i.e.* the mel spectrogram of five successive frames of audio. The reconstruction loss is used as the loss function of training process.

## 3. TRAINING STOP CRITERION

Unlike the case of training a well-performed variational auto-encoder, the key factor contributing to a successful submission in this task is to distinguish the in-domain samples with the outliers rather than reconstruct the inputs precisely. As a result, the stopping criterion of training process is expected to be different from those used for training a variational auto-encoder.

As the distribution of latent vectors follows Gaussian distribution, similar mel spectrogram data is expected to be represented by a similar latent vectors. As a result, the mel spectrogram of normal sound frames is expected to have a higher model likelihood regarding to the resulting Gaussian model in the latent space. For the mel spectrogram of anomaly sound frames, the model likelihood is expected to be lower. Using $\mathcal{L}(\mathbf{D}|\mathcal{M})$ to represent the likelihood in a per sample basis of dataset $\mathbf{D}$ for a given model $\mathcal{M}$, the anomaly score used in the proposed system is

$$s = \mathcal{L}(\mathbf{D}_{training}|\mathcal{M}) - \mathcal{L}(\mathbf{D}_{samples}|\mathcal{M}). \qquad (1)$$

The anomaly score $s$ is then used to calculate AUC and pAUC as required by the task.

Stated by Koizumi et al. [2] and Purohit et al. [3] the cause of anomaly machine hints that it is possible to have audio segments that are similar with normal machine audio in the anomaly sound of machines hence not all mel spectrogram in five successive frames in anomaly audio is necessary to be different from those of normal audio. As a result, the standard deviation of model likelihood for mel spectrogram of each five successive frames in a piece of audio (in the context *standard deviation of model likelihood*, SDML in short) is used as the indicator of the performance of resulting system, where a higher SDML may indicate a better performance of the resulting system.

The decision of setting a stopping point for SDML is not as easy as setting a threshold since the change of SDML is complicated over the training process. As using a whole dataset to train the proposed system leads to a faster converge in terms of the number epochs, a mini dataset that contains 51 pieces of audio as the training dataset and contains 51 pieces of both normal and abnormal pieces of audio as the testing dataset is used. With fewer samples in the dataset, the training loss converges much slower in terms of epoch, which enables a better observation on the change of SDML at each epoch as shown in Figure 1.

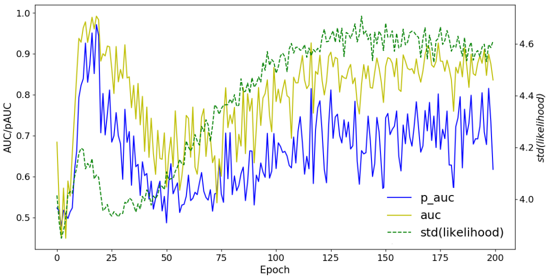

Figure 1: SDML changes at each epoch with a mini dataset used. The performance of resulting system where SDML is used as the indication of system performance. AUC and pAUC are used to evaluate the performance of the resulting system as required by the task.

The changes of SDML in terms of epoch have two extreme maximum points. The first extreme maximum point of SDML partially reflects to the best performance of system whereas the second extreme maximum point (and potentially the maximum point) of SDML fails to predict the system performance. As a result, the system proposes to use the first extreme maximum point as the training stop point. The resulting system after the training process is stopped is used as the candidate model for submission.

## 4. RESULTS

As indicated by the proposed SDML, the model likelihood of samples for the resulted Gaussian model in the latent space is used to calculate the AUC and pAUC of the resulting system as required by the task.

The performance of the proposed system in the development dataset has been shown in the following table. In Table 1, the performance of baseline system is compared with the proposed systems that trained by mini dataset and the whole training dataset. From the results, the proposed system has achieved a better performance significantly ($p = 0.0437$ for AUC and $p = 0.0012$ for pAUC with a significant level of 0.05) with the whole datasets ($\mathcal{D}_{whole}$) used.

Moreover, the system trained with a random selected mini dataset ($\mathcal{D}_{mini}$) also shows a better performance in terms of pAUC ($p = 0.0070$, significant level is 0.05) and has a comparable performance in terms of AUC (the hypothesis test fails to reject the null hypothesis $h_0$, two systems have unequal mean, at significant level of 0.05). Considering the fact that there are only 51 samples in the training dataset, the proposed algorithm with $\mathcal{D}_{mini}$ is much more data efficient compared with the baseline system.

The comparison of results between the proposed system and the baseline system suggests that a well-trained auto-encoder may not be able to detect the outliers as well as

| Machine | Baseline System | | Proposed ($\mathcal{D}_{whole}$) | | Proposed ($\mathcal{D}_{mini}$) | |
|---|---|---|---|---|---|---|
| | AUC | pAUC | AUC | pAUC | AUC | pAUC |
| fan00 | 54.41% | 49.37% | 89.05% | 73.45% | 5.47% | 47.36% |
| fan02 | 73.40% | 54.81% | 91.14% | 78.84% | 99.05% | 96.07% |
| fan04 | 61.61% | 53.26% | 61.61% | 52.25% | 32.58% | 48.80% |
| fan06 | 73.92% | 52.35% | 32.36% | 47.83% | 99.29% | 97.46% |
| pump00 | 67.15% | 56.74% | 100% | 100% | 100% | 100% |
| pump02 | 61.53% | 58.10% | 100% | 100% | 100% | 100% |
| pump04 | 88.33% | 67.10% | 99.98% | 99.89% | 100% | 100% |
| pump06 | 74.55% | 58.02% | 100% | 100% | 100% | 100% |
| slider00 | 96.19% | 81.44% | 100% | 100% | 100% | 100% |
| slider02 | 78.97% | 63.68% | 100% | 100% | 1.14% | 47.36% |
| slider04 | 94.30% | 71.98% | 80.12% | 63.48% | 84.97% | 73.00% |
| slider06 | 69.59% | 49.02% | 38.39% | 50.08% | 11.13% | 47.36% |
| ToyCar01 | 81.36% | 68.40% | 94.23% | 86.58% | 94.98% | 85.52% |
| ToyCar02 | 85.97% | 77.72% | 93.97% | 80.95% | 85.50% | 69.79% |
| ToyCar03 | 63.30% | 55.21% | 70.06% | 56.11% | 55.94% | 50.49% |
| ToyCar04 | 84.45% | 68.97% | 74.49% | 61.99% | 89.82% | 74.45% |
| ToyConveyor01 | 78.07% | 64.25% | 87.94% | 68% | 66.70% | 54.72% |
| ToyConveyor02 | 64.16% | 56.01% | 77.89% | 57.86% | 61.48% | 52.26% |
| ToyConveyor03 | 75.35% | 61.03% | 63.72% | 53.74% | 91.72% | 75.24% |
| valve00 | 68.76% | 51.70% | 80.71% | 58.77% | 87.01% | 65.45% |
| valve02 | 68.18% | 51.83% | 99.71% | 98.50% | 3.95% | 47.36% |
| valve04 | 74.30% | 51.97% | 87.03% | 74.38% | 91.06% | 79.56% |
| valve06 | 53.90% | 48.43% | 61.29% | 54.16% | 53.50% | 50.61% |

Table 1: Performance of the proposed system in terms of AUC and pAUC. A higher value indicates better performance.

the an auto-encoder that has not been trained to converge (named as "semi auto-encoder" in this report). In other words, a well trained variational auto-encoder may have an ability of convert an outlier sample to a normal sample hence a semi auto-encoder may outperform a well-trained auto-encoder in a certain range of tasks.

## 5. FUTURE WORKS

Although the proposed method has obtained success on detection anomaly sound for some extent, there any many research problems has yet to be solved. For example, it is still hard to explain why there are two extreme maximum points in terms of epochs for the SDML as explaining such explanation requires a better insight of how a variational auto-encoder is trained. Moreover, from the results of development datasets, the best performed system usually come after few epochs compared with the proposed selection of stopping points of training. Further investigations should be expected to find a better stop training point. Besides SDML and model likelihood in latent space, it is possible to find a better pairing of the training stopping criterion and the performance scoring variables.

## 6. CONCLUSION

With the architecture of variational auto-encoder, a semi variational auto-encoder is trained, which stops training process according to the standard deviation of model likelihood of five successive audio frames in the resulting latent space. The resulting system then use the model likelihood to calculate the anomaly score. The results show that the proposed system achieves a better performance than the baseline system.

## 7. REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," stat, vol. 1050, p. 1, 2014.

[2] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "Toyadmos: A dataset of miniature-machine operating sounds for anomalous sound detection." IEEE, 2019, pp. 313–317.

[3] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), Nov. 2019, pp. 209–213.