ACOUSTIC ANOMALY DETECTION BASED ON SIMILARITY ANALYSIS

Zhao Shuyang

Tampere University, Finland.

ABSTRACT

This study uses nearest neighbour distance as a measure of anomaly. The nearest neighbour distance is defined as the distance from a test sample to its nearest neighbour in the training dataset, which contains only sounds recorded in normal condition. A sample is represented by a multi-variate Gaussian distribution of corresponding MFCCs. Kullback-Leibler divergence is used to measure the dissimilarity between two distributions, and it is further used as a distance between two samples. Three submissions vary in the use of MFCC deltas and the type of covariance matrices used for Gaussian distributions.

Index Terms: Anomaly detection, nearest neighbour, KL divergence

1. INTRODUCTION

Anomalous sound detection aims at identifying anomaly in target machine based on the sounds emitted. The manual monitoring of industrial machines could be time-consuming. Computational audio content analysis enables almost effortless monitoring of industrial machines. In addition, the accuracy of the computational analysis might be better than human, since human auditory system is not well trained to perform this type of task.

Previously, several studies have been made using autoencoders to perform anomalous sound detection such as [1]. However, the evaluation on these methods have been limited to small scale datasets. Lately two datasets [2, 3] are released specifically for the purpose of anomalous sound detection study. These datasets enables evaluation on real industrial machine monitoring scenarios.

In this study, the anomaly detection problem is understood as a variation of retrieval problem, whether a similar recording can be found in the corpus of normal condition recordings. A recording is considered as anomaly, if no similar recording can be found. To the best of our knowledge: three methods have been used to measure the similarities between audio pairs: KL divergence between feature distributions, dynamic time warping (DTW) between feature sequences and cosine similarities between embedding vectors. In order to choose the proper similarity metric, a manual analysis has been made on the datasets. As the first observation, the recordings are rather long (10 seconds) and the feature sequence order is not important to most of the device types. Due to the large computation cost, DTW is excluded in this study. As the second observation, the anomaly can hardly be detected by human, who has no prior knowledge about the devices. Based on this observation, the little general knowledge learned from large datasets such as Audioset can be transferred to this problem. As a negative effect, the embeddings extracted with a model learned from general sound event detection dataset may discard information that is important to the device anomaly but irrelevant to sound event detection. As a summary of the analysis, KL divergence between feature distributions is the most suitable choice as the similarity metric.

2. THE PROPOSED METHOD

The proposed method measures the dissimilarity, or distance, between each test sample and training sample pair. The anomaly score of a test sample is determined by its distance to the nearest sample in the training set. The dissimilarity between two sounds is measured based on KL divergence between the mel-frequency cepstral coefficients (MFCCs) distribution in the two sounds.

2.1. MFCC extraction

The extraction of MFCCs is as follows. The frame length is 200 ms and the hop length is 50 ms. Given the sampling rate is 16 000 Hz, the frame length and hop length is 3200 samples and 800 samples, respectively. A relative long frame length is used, since sounds produced by the industrial machines change slowly over time. The number of mel-bands is 256. MFCCs of 40 coefficients are used besides the first and second order deltas. The deltas of a frame is computed based on the next four frames and past four frames. Both the MFCCs and their deltas are extracted using librosa.

2.2. Sound segment representation

A sound segment, or an audio file in a dataset, is represented by a multi-variate Gaussian distribution as $\mathbf{P}_i = \mathcal{N}(\mu_i, \Sigma_i)$, based on the mean and covariance of the MFCCs in the segment. The MFCCs within a segment i of n frames are denoted as $\mathbf{X}_i = \mathbf{x}_{i,1}, \mathbf{x}_{i,2}, ...\mathbf{x}_{i,n}$. The mean of MFCCs is denoted as

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n} \mathbf{x}_{i,j}}{n}. \tag{1}$$

Two different types of covariance matrices are used in submissions. The first one uses diagonal covariance matrices, using the variance of each variable as diagonal values, computed as

$$\boldsymbol{\Sigma}_i = diag(\boldsymbol{\sigma}_i) = diag(\frac{\sum_{j=1}^{n}(\mathbf{x}_{i,j} - \boldsymbol{\mu}_i)^2}{n}). \tag{2}$$

This assumes the variables independent to each other. The other one uses directly the covariance matrix as

$$\boldsymbol{\Sigma}_i = \frac{(\mathbf{X}_i - \boldsymbol{\mu}_i)(\mathbf{X}_i - \boldsymbol{\mu}_i)^T}{n}. \tag{3}$$

2.3. Segment-to-segment dissimilarity measurement

KL divergence is a measurement of dissimilarity between two distributions. In this work, it is used to determine the distance between a a sound segment pair. The sound similarity metric has been used in sound information retrieval [4] and clustering-based active learing [5].

The KL divergence between two multi-variate Gaussian distributions $\mathcal{P}_0$ and $\mathcal{P}_1$ is calculated as

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{P}_0\|\mathcal{P}_1) =& \frac{1}{2}(\mathrm{tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0) \\ &+ (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) \\ &+ \ln(\frac{\det\boldsymbol{\Sigma}_1}{\det\boldsymbol{\Sigma}_0}) - k), \end{aligned} \tag{4}$$

where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are mean and covariance of distribution $\mathcal{P}_0$, respectively. The mean and covariance of distribution $\mathcal{P}_1$ are denoted as $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$.

KL divergence is not a commutative operation so that $D_{\mathrm{KL}}(\mathcal{P}_0\|\mathcal{P}_1)$ is different from $D_{\mathrm{KL}}(\mathcal{P}_1\|\mathcal{P}_0)$. In order to obtain a symmetric dissimilarity matrix, the average of both way KL divergence is used to measure the dissimilarity between two segments i and j as

$$d_{i,j} = \frac{D_{\mathrm{KL}}(\mathcal{P}_i\|\mathcal{P}_j) + D_{\mathrm{KL}}(\mathcal{P}_i\|\mathcal{P}_j)}{2}. \tag{5}$$

2.4. Variation in submissions

The submission 1 as the default setup uses the same setup for all the devices. It uses MFCCs, deltas and second order deltas. The Gaussian distribution is based on diagonal covariance matrices.

In submission 2 and submission 3, the devices are divided into two groups: stationary sound machines, which produce mostly stationary sound; transient sound machines, which produce important signature transient sounds. The stationary sound machine group includes fan, pump, toy car and toy conveyor. The transient sound machine group includes slider and valve. The deltas are more likely to be affected by environmental noise, compared to stationary sound machines. Thus the deltas are excluded in submission 2 and submission 3 for stationary sound machines. Diagonal covariance matrix assums independence in variables. MFCCs are decorrelated with discrete time transform. However, the distribution of deltas could be dependent to the static MFCCs. Taking this into account, full covariance matrices are used for the transient sound machines in submission 3.

2.5. Anomaly score

The distances based on KL divergence are measured between each test sample and each training sample. The nearest distance from a test sample to any of the training samples is used as the anomaly score of the test sample. The anomaly score of a test sample i is thus computed as

$$s_i = max(d_{i,j}|j \in \mathcal{T}), \tag{6}$$

where $\mathcal{T}$ is the trainind set.

3. EVALUATION

The evaluation in this technical report is based on only the development dataset, since the evaluation set is not released by the time of writting.

3.1. Experimental setup

The three submissions use the same basic setups following the methods described in the method section. Three variables are tested, the number of MFCCs, the use of MFCC deltas and the type of covariance matrices. The setups in the three submissions are shown in Tables 1. Notably, the first submission uses exactly the same setup for all the machine types.

	Sub. 1	Sub. 2	Sub. 3
Deltas(stationary)	Yes	No	No
Deltas(transient)	Yes	Yes	Yes
Covariance(stationary)	Diagonal	Diagonal	Diagonal
Covariance(transient)	Diagonal	Diagonal	Full

Table 1. Setups in the three submissions. Fan, ToyCar, ToyConveyor and pump belong to the stationary group. Slider and valve belong to the transient group.

3.2. Results

The experimental results on each machine type is shown in Tables 2. Comparing the results on stationary sound machines between submission 1 and submission 2, it is slightly better to use only static MFCCs for stationary sound devices except pump. In the manual analysis, pump seems to produce some transient sounds, but it does not have a clear pattern. From this perspective, it is hard to determine which group it belongs. Comparing the results on transient sound machines between submission 1 and submission 3, it is clearly better to use full covariance Gaussian distribution for MFCCs and their deltas.

4. REFERENCES

[1] Erik Marchi, Fabio Vesperini, Florian Eyben, Stefano Squartini, and Björn W. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*. 2015, pp. 1996–2000, IEEE.

[2] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Noboru Harada, and Keisuke Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, November 2019, pp. 308–312.

[3] Harsh Purohit, Ryo Tanabe, Takeshi Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei 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)*, November 2019, pp. 209–213.

[4] Yichi Zhang and Zhiyao Duan, "IMISOUND: an unsupervised system for sound query by vocal imitation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. 2016, pp. 2269–2273, IEEE.

[5] Zhao S.Y., Toni Heittola, and Tuomas Virtanen, "Active learning for sound event classification by clustering unlabeled data," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, 2017, pp. 751–755.

Machine ID	Submission 1	Submission 2	Submission 3
Fan 0	0.663	0.682	-
Fan 2	0.896	0.900	-
Fan 4	0.787	0.796	-
Fan 6	0.883	0.892	-
Fan average	0.807	**0.817**	**-**
Pump 0	0.829	0.817	-
Pump 2	0.818	0.813	-
Pump 4	0.972	0.971	-
Pump 6	0.840	0.837	-
Pump average	**0.865**	0.858	-
ToyCar 1	0.854	0.882	-
ToyCar 2	0.946	0.962	-
ToyCar 3	0.791	0.829	-
ToyCar 4	0.982	0.992	-
ToyCar average	0.893	**0.916**	**-**
ToyConveyor 1	0.784	0.818	-
ToyConveyor 2	0.674	0.707	-
ToyConveyor 3	0.788	0.806	-
ToyConveryor average	0.749	**0.777**	**-**
Slider 0	0.949	-	0.975
Slider 2	0.836	-	0.828
Slider 4	0.926	-	0.986
Slider 6	0.715	-	0.922
Slider average	0.856	-	**0.928**
Valve 0	0.877	-	0.909
Valve 2	0.815	-	0.965
Valve 4	0.851	-	0.977
Valve 6	0.719	-	0.802
Valve average	0.815	-	**0.913**

Table 2. AUC for machines. The symbol - is used when the method is the same as previous submission and the result is the same as the left column.