A METHOD OF ANOMALOUS SOUND DETECTION WITH MULTI-DIMENSIONAL AUDIO FEATURE INPUTS

Youfang Han, Dinghou Lin , Chunping Li

Tsinghua University
School of Software
Beijing, China
lindh19@mails.tsinghua.edu.cn

Ran Shao

ELCOM (Suzhou) Co. Ltd.
Suzhou, China
shao.ran@elcom.cn

ABSTRACT

In this technical report, we describe the system we submitted to DCASE2020 task 2 in details, i.e., anomalous sound detection (ASD). The goal is to train a model which can distinguish normal sound and abnormal one when only normal sound samples are used as training data. To achieve this goal, we need to find out the characteristics of normal sound. Firstly, we adopt the preprocessing method to intensify the features of normal audio, secondly, we extract different types of features including artificial features and implicit features. Moreover we also use psychoacoustic features to assist to train the model. Finally, we achieve better performance than DCASE2020 baseline system.

Index Terms—preprocessing, feature extraction, artificial features, psychoacoustic features, implicit features

1. INTRODUCTION

Anomalous sound detection (ASD) is the task to identify whether the sound emitted from a target device is normal or anomalous. The main challenge of this task is to detect unknown anomalous sounds under the condition that only normal sound samples have been provided as training data collected from the real-world factories, and actual anomalous sounds rarely occur and are highly diverse. Therefore, exhaustive patterns of anomalous sounds are impossible to deliberately make and/or to collect. This means that we have to detect unknown anomalous sounds which were not observed in the given training data. This task cannot be solved as a simple classification problem, even though the anomaly detection problem seems to be a two-class classification problem.

2. PROPOSED METHOD

This work aims at the data set of task 2 and detects abnormal audio. The basic technical steps are shown in Fig.1 including audio preprocessing, feature extraction, modeling analysis and the final result output refereed to [1].

Fig 1. basic technical steps

In the preprocessing stage, we filter the audio frequency and divide into different sub-frequency bands. Since the frequency bands of different audio's main features are different, we adopt different filtering methods to process different data sets. In the feature extraction stage, the log Mel energies data of audio is used as main input features for model training. At the same time, we also introduce statistical methods to analyze the spectogram and the waveform to find out some possible abnormal sound patterns. In addition, some psychoacoustic parameters are used in the training of the model. In the stage of audio modeling and analysis, the AutoEncoder is built based on the baseline method, and the extracted features are analyzed. Finally, the mean square error method is used to produce the score of abnormal value.

2.1. Preprocessing

The audio in DCASE Test2 dataset is mixed with a lot of disturbing noise data that leads to the difficulty to distinguish the differences between normal and abnormal sounds. The frequency range of the background noise is different from the target frequency range. Therefore, it is possible to filter out some unrelated background noise frequency band, therefore, the better SNR (Signal Noise Ratio) can be achieved.

In the case of dataset “slider”, because of the background noise, it is difficult to intuitively hear the difference between normal sound and abnormal sound. After the slider's audio is filtered, the characteristic difference between normal and abnormal audio can be clearly heard by only observing one or two frequency ranges.

2.2. Feature extraction

Human auditory system can easily distinguish normal and abnormal audios, but how to characterize the difference between normal and abnormal ones is a challenging problem. Audio signals can be described in many ways, such as MFCC, log Mel energies, spectrogram, waveform and so on. Taking spectogram as an example, when we do short-time Fourier transform for audio, we can intuitively see the change rule of audio frequency with time. By observing spectogram of normal and abnormal audios, we can intuitively find some features to distinguish normal and abnormal audios.

In our experiment, based on the baseline method, three kinds of audio feature extraction methods are proposed to optimize the model, i.e., artificial features, implicit features and psychoacoustic features. Both of the artificial and implicit

features are extracted by the spectogram and waveform representations of audios.

2.2.1. Artificial feature construction

For specific kinds of audio data, we can construct some features as the normal and abnormal partition indexes by analyzing the spectogram and raw waveform of audios. This partition method can be applied to some special sound patterns [5], such as squeak, rattle, and tap value, etc. Through these values the strength of squeak pattern, rate pattern and tap pattern can be described respectively in a piece of audio. The higher the value, the higher the intensity of the sound pattern in the whole audio. This kind of dividing index is the characteristic that we define artificial features

For example, when listening to the three data sets of pump, slider and valve, as well as the electric-engine in the extended data set, we can hear the "tap-tap" like knocking sound in the abnormal audio. This kind of sound mode can be used as a characteristic index to judge whether the audio is abnormal. By analyzing the spectogram and raw waveform data from audio with or without "tap-tap" like knocking sound, we can find the difference between them in both frequency domain and time domain. In view of this difference, a specific algorithm is designed by using statistical methods, and a set of vectors can be obtained to represent the severity of the "tap" features in an audio segment. This set of vectors is used as artificial features to assist model training. Through the introduction of artificial features, it has a good experimental result on the three data sets of pump, slider and valve in DCASE2020 task2.

In addition, there are various types of artificial features that can be used in audio quality detection experiments, which have unexpected effects on different types of audio data.

2.2.2. Implicit feature extraction

By observing the spectogram and raw waveform of normal and abnormal audios, some differences can be found and be defined intuitively. However, some other features, which are not obvious and even cannot be detected by human auditory system, here are called as implicit" features.

Deep neural network is a good method to find "implicit" features in high-dimensional data [6]. Some embedded representations or patterns in the raw data can be selected out after the training in deep neural networks. This method is used for both 2D spectogram and 1D waveform. These "implicit" features are hard to interpret and cannot be defined by the artificial formulas, but it may play an important role in feature extraction.

In DCASE task2, two external datasets VGGish and OpenL3 are provided. These datasets can be used to extract the embedding representation of the audios.

2.2.3. Psychoacoustic parameter

Psychoacoustics is the study of the relationship between subjective auditory perception and objective physical quantities of sound. In order to find out the relationship between sound and human sensation, it is necessary to correlate the physical parameters of audio signals, such as sound pressure level, frequency, modulation and the parameters related to hearing [2].

For the data set of DCASE task2, there are some differences in auditory perception between abnormal audio and normal audios. For example, the audio in slider and pump is noisy, while the audio in fan dataset is sharp. Therefore, we introduce four psychoacoustic parameters, i.e., loudness [3], sharpness, roughness [4] and brightness to assist model training.

2.3. Model design

The designed model is mainly composed of three parts. Firstly, preprocessing and feature extraction are conducted at the input, and then training is carried out in autoencoder. Finally, the mean square error method is used to produce the scores of outliers and calculate AUC and pAUC to evaluate the detection effect.

The input data is composed of four parts, including the log-mel energies data of the audio used by the baseline method, as well as inputting artificial feature vectors, psychoacoustic parameters, and implicit feature vectors extracted through neural networks to participate in model training. The artificial feature vectors and implicit features are obtained by analyzing the raw waveform and spectogram of audio.

The core discriminant model is the AutoEncoder. AutoEncoder is modeled based on the baseline method and consists of Encoder and Decoder. The loss function uses the mean square error. If the input data has a high degree of restoration through this process, there is a greater probability that it is normal audio. The model diagram is as follows.

The designed model has strong scalability. When a new "artificial features is constructed, it can be added to the model. At the same time, audio representations such as log-mel energie and MFCC can also be embedded to extract the implicit features, and then used as the input of the model.

3. EXPERIMENT AND RESULTS

This section describes the experiments and results on the development dataset and evaluation dataset.

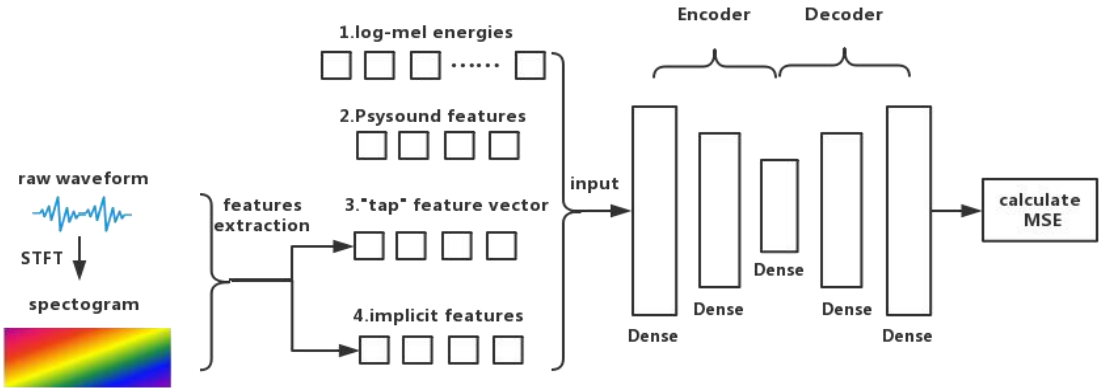

Fig 2. the structure of the model

3.1. Dataset

The experimental data set uses the development dataset and evaluation dataset provided by DCASE task2. There are six types of audio data, i.e., fan, pump, slider, valve, ToyCar and ToyConveyor, which represent different types of sound. Different numbers in each dataset represent different types of noise interference. In the development dataset, the training set only contains normal audio samples, while the test set contains positive and negative audio samples. In Evaluation dataset, the training set contains normal audio samples, and the test set data is unlabeled.

3.2. Experimental setup

This experiment is carried out using Python 3.6, the deep learning framework TensorFlow-gpu and third-party libraries for audio processing including librosa, vamp, and timbral-models, etc. AutoEncoder is built using Python's keras library, and in the network the mean square error is used as loss function, and the Adam used as optimizer.

3.3. Results

According to the model mentioned before, experiment with six types of audio data in the development dataset and evaluation dataset respectively.

3.3.1. Results of development dataset

Use Baseline's method and our method to train the training set in the development dataset, apply the trained models to the test set, and calculate the AUC and pAUC values.

The experimental results of the development data set are shown in Table 1 and Table 2.

Table1 Experimental results of baseline method

Dataset	averaged_auc	averaged_pauc
ToyCar	78.77	67.58
ToyConveyor	72.53	60.43
Fan	65.83	52.45
pump	72.89	59.99
slider	84.76	66.53
Valve	66.28	50.98

Table2 Experimental results of our method

Dataset	averaged_auc	averaged_pauc
ToyCar	79.01	66.32
ToyConveyor	74.95	62.50
fan	66.22	52.43
pump	74.77	61.04
slider	91.56	82.24
valve	81.55	55.72

Comparing our method with the Baseline method, the averaged_auc of the six data sets has been improved. The slider and valve data sets have the most obvious effect. The fan data set has the worst effect. We listened to most of the audio in the fan data set, and it is difficult to make a clear distinction between normal and abnormal audio. On the contrary, the audio in the slider and valve data sets can clearly hear the different sound patterns in the abnormal audio than the normal one. The model can fit the normal audio data well, and the experimental effect is better. For ToyCar, ToyCoveyor and pump dataset, our method is not much different from the Baseline method, probably because the sound feature pattern we found does not apply to these data sets.

3.3.2. Results of evaluation dataset

We use the proposed method to train the model with evaluation dataset, and then apply the model to the test set. The mean square error method is chosen to calculate the abnormal score. Since the test set has no public labels, it cannot calculate averaged_auc and averaged_pauc.

Detailed outlier information is given in the attachment of the experimental results. Here, the average, maximum and minimum values of the outlier scores of the six sets of audio data are shown in Table 3.

Table3 Experimental results of evaluation dataset

Dataset	averaged_ score	maximum_ score	minimum_ score
ToyCar	10.25	13.05	9.34
ToyConveyor	10.22	14.21	8.99
Fan	10.58	35.12	8.13
Pump	10.54	26.92	7.29
Slider	9.74	12.15	8.63
Valve	9.08	11.93	8.67

Slider and valve have better effects in the experiment of development dataset. Due to the obvious difference between its normal and abnormal audio in characteristics, the overall abnormal value is low. Fans and pumps dataset, it is difficult to distinguish the difference between normal and abnormal audio, the model is also more difficult to fit, and the overall abnormal value is higher.

4. CONCLUSION

Overall, the results of the most datasets are improved in comparison to the baseline result, but some of them have minor changes. For example, the effect of fan dataset is not obvious. The methods applied in this experiment are more applicable to solve specific problems. In the future, we plan to find he more general method that can be applied to various types of anomalies.

5. REFERENCE

[1] Francesc Alías, Joan Socoró, Xavier Sevillano. A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds. 2016, 6(5)

[2] Groen J J, Versteegh R M. Frequency Modulation and The Human Ear[J]. Acta Oto Laryngologica, 1957, 47(5):421-430.

[3] ISO 532-1:2017. Acoustics − Methods for calculating loudness−Part 1: Zwicker method[S].

[4] Robinson D.W.. Psychoacoustics: facts and models: 1990, by E. Zwicker and L. Fastl. Berlin-Heidelberg-New York: Springer-Verlag. Price (hard cover) DM98Â· 00; pp. 354 + x; 274 figs. ISBN 3-540-52600-5[J]. Academic Press,1991,149(3).

[5] G.-J. Lee, J. Kim, K. Kim. An algorithm to automatically detect and distinguish squeaks and rattles[J]. sound & vibration, 2015, 49(9):8-10.

[6] Piczak K J . Environmental sound classification with convolutional neural networks[C]// 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2015.

[7]Koizumi Y , Saito S , Uematsu H , et al. ToyADMOS: A Dataset of Miniature-Machine Operating Sounds for Anomalous Sound Detection[J]. 2019.

[8]Purohit H , Tanabe R , Ichige K , et al. MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection[J]. 2019.

[9]Koizumi Y , Kawaguchi Y , Imoto K , et al. Description and Discussion on DCASE2020 Challenge Task2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring[J]. 2020.