SUBMISSION TEMPLATE FOR TISMIR

A TEMPORAL APPROACH TO ANOMALOUS SOUND DETECTION:

Ashlae Blum, GEM Fellow, August 12, 2021

Keywords: music information retrieval, anomalous sound detection, librosa, percussive density, unsupervised learning

1 INTRODUCTION

The problem of unsupervised anomalous sound detection (ASD) remains a complex endeavor in the global AI community. In contrast with its supervised or semi-supervised counterparts, unsupervised learning does not require labeled data to train the neural network. The model determines which data are normal and which are anomalous from a set of unlabeled data [1]. This strategy presents some obvious advantages. In a physical sense, models can be deployed in areas unreachable by or intolerable to humans, and physical maintenance or labor costs can be minimized. From a scientific perspective, unexpected or otherwise unknowable patterns in the data may be observed, which can lead to innovations or breakthroughs in research. A notable downside of this method is that there is an accuracy tradeoff as compared with semi-supervised or supervised learning [16]. In cases where ML models are deployed on edge devices, such as in some industrial settings, latency, power limitations and therefore complexity are also a consideration [2].

Predictive Maintenance (PM) is one such application of ASD. PM employs anomaly detection and classification to predict abnormal behavior and then determine the cause [3, 4]. For example, industrial settings are both noisy and can also be full of unexpected or sudden sounds. Some particular audio event might not necessarily be that of the machine breaking. In this case, it may be more useful to determine whether the quality of the audio has changed rather than the detection of a single incident. This is known as scene, rather than binary, classification. An application of this might be to use audio sensors to capture real-time data, and then analyze it for anomalies. In the case of high energy physics facilities such as at Fermilab, this method could be used to detect and prevent a superconducting magnet quench from happening [5]. Benefits of this include preventing danger to humans, and conserving physical as well as financial resources.

2 MOTIVATION

When we think about operating everyday appliances, we may think about lawnmowers, blenders, or vehicles, and the characteristic sounds of their electrical components. These appliances tend to have a particular frequency and vibrational pattern for their normal operation. With vehicles, then, the sounds of the starter, ignition, or engine idling can be clear indicators of abnormal operation. Even if there is no single noise that stands out as before-after or cause-effect, our ears will be finely attuned to the quality of sound that may tell us something is off.

To-date, much of the existing literature with respect to ASD tends to focus on spectral representations and frequency analysis [9]. However, we believe there is unexplored terrain with respect to temporal analysis, also known as onset detection. Two existing studies that utilize temporal features to determine road hazards include [3] and [9], however they are sparse in their extensibility to our inquiry in that they focus on single event detection rather than scene classification.

Thus, without making prior assumptions as to the results we hope to obtain, we seek to engage in an experimental approach to understanding the role of *percussive density** in determining the sound quality of audio. Our hope is to define and compare a set of temporal characteristics such that a correlation between normal and abnormal audio may be determined.

* *percussive density* is a term coined by the author to denote some metric of audio onsets

3 BACKGROUND

In this section, we present a variety of audio features that are useful in MIR and ASD. These features are implemented in Librosa [8], and we colloquially define them here for context and future reference.

3.1 Spectral Features of Audio

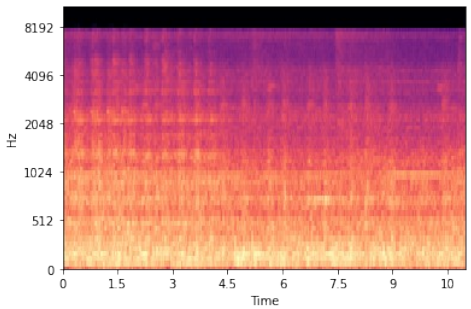

Figure 1. Power Spectrogram

Short Time Fourier Transform (STFT): Fourier transform that determines local phase and frequency of a signal

Power Spectrogram: Signal strength across frequencies of a time varying audio signal

Mel Frequency Cepstral Coefficients (MFCC): Fourier coefficients of the short term power spectrum of an audio signal scaled to the nonlinear human auditory response

Chroma: Pitch class representations to encode harmony while minimizing while suppressing variations in octave, loudness, and timbre.

3.2 Temporal Features of Audio

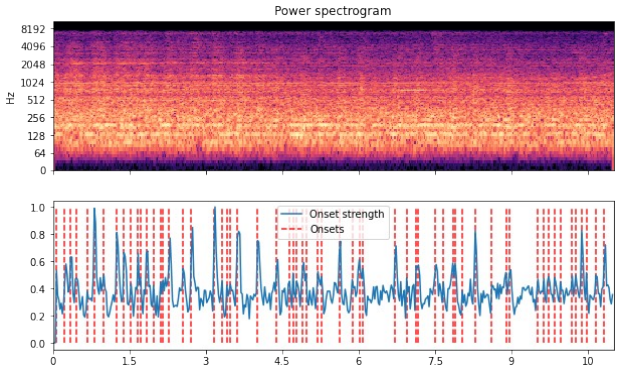

Figure 2. Spectrogram and Onset Detection

Onset Strength: Representation of spectral flux across audio frames

Onset Detection: Peak positions of the onset curve

Tempo: estimates the average global tempo in beats per minute (bpm)

PLP (predominant local pulse): Positive normalized local beat tracking as determined by tempogram analysis of onset events

4 APPROACH

We begin this inquiry by implementing the baseline model presented in the DCASE2020 Task2 benchmark challenge [6]. Our dataset comprises industrial audio in the form of approximately 7000 *.wav* files from the ToyADMOS dataset, specifically, that of the Toy Car [7]. We train this model on a Windows10 64-bit operating system. Anomaly scores are calculated for each audio file, and are then plotted to visualize the accuracy of the model.

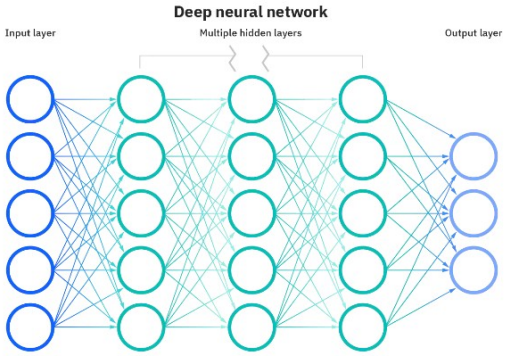

Figure 3. Deep neural network

Using techniques informed by a familiarity with Music Information Retrieval (MIR), we next explore low-level audio features of the raw data using the Librosa package for Python [8]. Datasets of spectral and temporal information are constructed using these methods. We determine that, on detour from a typical ASD emphasis towards examining spectral information, there may be merit in correlating temporal information from the audio. The bulk of the idea is to compare tempo and PLP data across audio files to see if correlations emerge. Upon closer inspection, we conclude that a Long Short Term Memory (LSTM) neural network will be the most appropriate way to engage in a temporal approach for this problem.

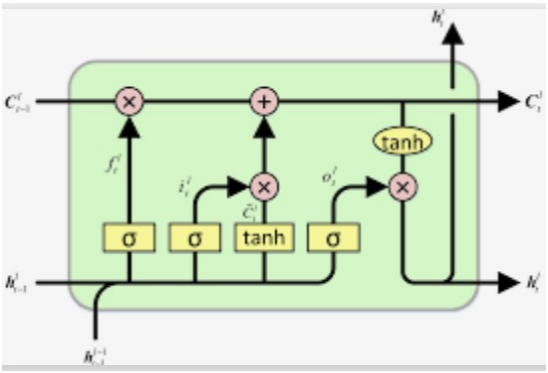

Figure 4. LSTM network

5 METHODS

5.1 Anomaly Score, Spectral Approach

We use an autoencoder to train on 7000 .wav files of normal audio data (each 10 seconds long). This allows us to determine an anomaly score for each audio file, from which the model's accuracy can be calculated. The autoencoder uses an Adam optimizer with a mean squared error loss function, and receives datasets of MFCCs as its input. Using a sliding window of 5 frames with 128-mel bins and a hop length of 512, the 9-layer neural network compares spectral data to determine self-

similarity and loss minimization over 100 epochs of training. The network structure is comprised of a 4x128, 1x8, 4x128 node architecture (270k parameters), and ReLU and BachNorm are implemented after each hidden layer [17].

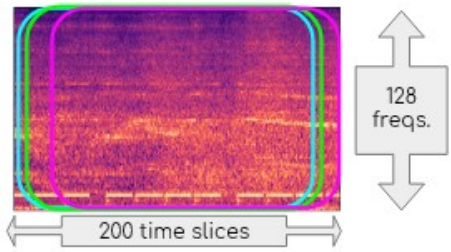

Figure 5. Sliding Window Method Spectrogram

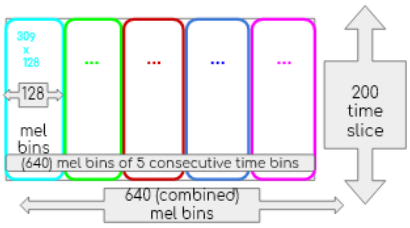

Figure 6. Sliding Window Method Frames

The resulting anomaly scores are determined by a threshold for the loss function, indicating the degree to which an audio file is anomalous or normal [15]. Receiver Operator Characteristic (ROC) curves, Area Under Curve (AUC) and Partial AUC (pAUC) are calculated as a measure of the model's accuracy.

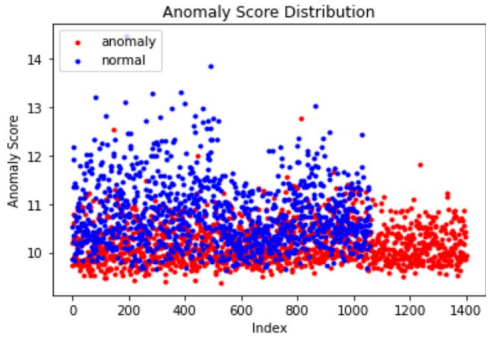

Figure 7. Scatterplot of Anomaly Scores

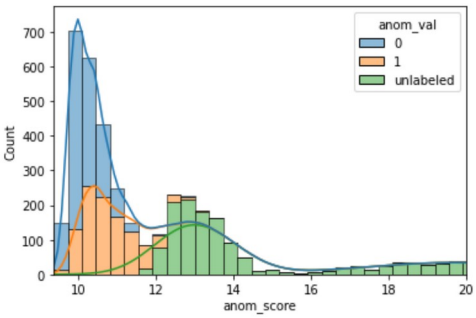

Figure 8. Histogram of Anomaly Scores

5.2 Temporal Approach

Through exploratory means using Librosa, we evaluate the audio data and compute the average tempo and PLP across each file. Tempo is a global average, while PLP is variably valued and non-uniformly distributed across each audio file.

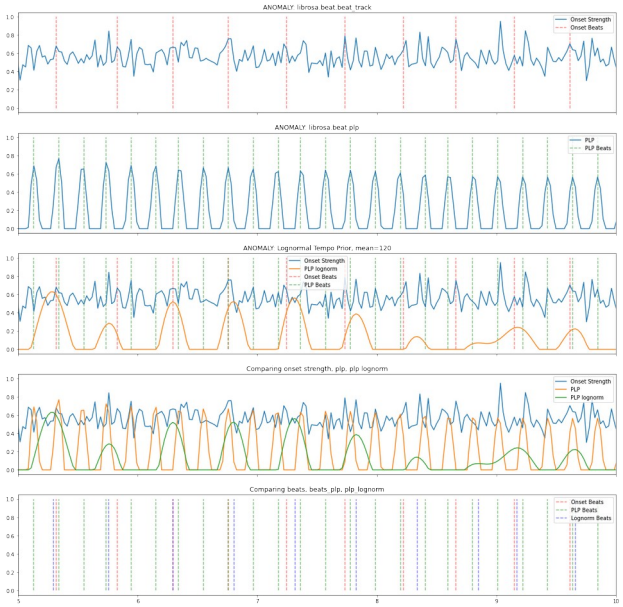

Figure 9. Temporal Features of Audio

6 RESULTS / CONCLUSIONS

6.1 Anomaly Score, Spectral Approach

```
ToyCar
id,AUC,pAUC
01,0.8032792207792209,0.6615402141717931
02,0.8513854447439353,0.747595403603348
03,0.6234932614555256,0.5417222301035608
04,0.833401617250674,0.6728046531422897
Average,0.7778898860573389,0.6559156252552479

```

Figure 10. AUC and pAUC scores

For the anomaly score calculation portion of this study, our model computes an AUC of 0.778. Using the standard AUC scale of 0.5 to 1.0, where 0.5 is 0% accuracy and 1.0 is 100% accuracy, this indicates a fair to good performance. While there are a variety of parameters that could yet be modified to improve accuracy, our goal for this portion of the experiment has been to derive our own benchmark such that we could train our own datasets of audio features on this model. For this purpose, our results are quite acceptable.

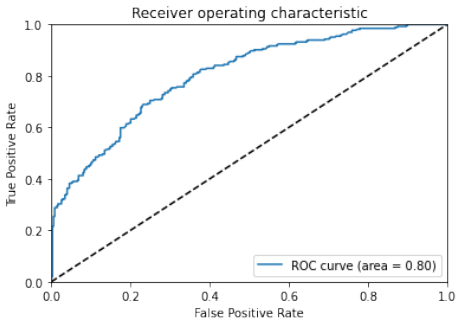

Figure 11. ROC for first data subset

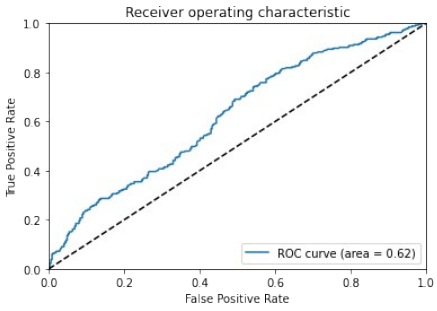

Figure 12. ROC for second data subset

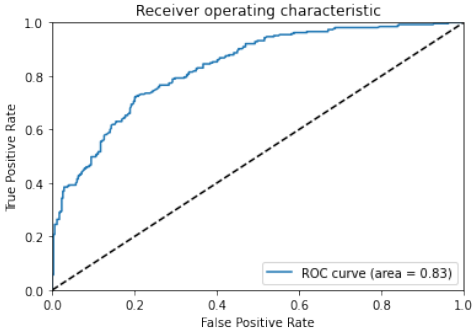

Figure 13. ROC for third data subset

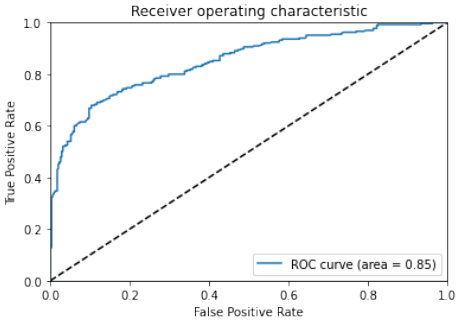

Figure 14. ROC for fourth data subset

6.2 Temporal Approach

The goal of this research is to compare the tempo and PLP data of each file to itself, as well as comparing this information across all files. One projected approach is to define a metric or scoring mechanism for the temporal data. For example, one value might indicate a topology, or self-similarity score that describes each unique audio file. Another value might compare how such scores are similar across the set of all audio files. However, in the process of this line of inquiry, we discovered that the size of the temporal data for each audio file turns out to be non-uniform. That is, different audio files have different numbers of datapoints, resulting in different lengths of values within the dataset. As such, we must construct a different type of model to receive the data, or, less ideally, determine an appropriate way to normalize the data manually. We posit that using a Long Short Term Memory network to train the data will be an appropriate way forward, since it is able to accept variably sized data.

7 FUTURE WORK

We plan to extend this work to correlate the occurrence of spectral and temporal features, and to make the resulting method portable for implementation on edge devices using TinyML [18]. The hope is that this application will be usable for deployment in industrial applications, such as to detect the quench of superconducting magnets.

8 ACKNOWLEDGEMENTS

We would like to extend our thanks to Ryan Rivera, Nhan Tran, Ben Woods, and the staff at Fermilab National Accelerator Laboratory and the GEM Institute for their support and encouragement on this project.

9 REFERENCES

[1] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Yuta Kawachi, Noboru Harada. Unsupervised Detection of Anomalous Sound Based on Deep Learning and the Neyman–Pearson Lemma. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 27, No. 1, January 2019, pp. 212.

[2] Fahim, F. hls4ml: An Open-Source Co-Design Workflow to Empower Scientific Low-Power Machine Learning Devices.

[3] D. Henze, K. Gorishti, B. Bruegge and J. Simen, "AudioForesight: A Process Model for Audio Predictive Maintenance in Industrial Environments," 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 2019, pp. 352-357, doi: 10.1109/ICMLA.2019.00066.

[4] Pech M, Vrchota J, Bednář J. Predictive Maintenance and Intelligent Sensors in Smart Factory: Review. Sensors (Basel). 2021;21(4):1470. Published 2021 Feb 20. doi:10.3390/s21041470

[5] D. Hoang et al., "Intelliquench: An Adaptive Machine Learning System for Detection of Superconducting Magnet Quenches," in IEEE Transactions on Applied Superconductivity, vol. 31, no. 5, pp. 1-5, Aug. 2021, Art no. 4900805, doi: 10.1109/TASC.2021.3058229.

[6] Y. Koizumi. Description and Discussion on DCASE 2020 Challenge Task2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring

[7] Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, ToyADMOS: A Dataset of Miniature-Machine Operating Sounds for Anomalous Sound Detection. Proc. of the Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA), 2019.

[8] Brian McFee, C. Raffel, D. Liang, Dan Ellis, M. McVicar, E. Battenberg, & O. Nieto. 2015. librosa: Audio and Music Signal Analysis in Python.

[9] Eduardo Carvalho Nunes. 2021. Anomalous Sound Detection with Machine Learning: A Systematic Review. January 2021.

[10] A. Yamashita, T. Hara, and T. Kaneko. 2006. “Inspection of visible and invisible features of objects with image and sound signal processing,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., 2006, pp. 3837–3842.

[11] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada. 2017. Optimizing acoustic feature extractor for anomalous sound detection based on Neyman–Pearson lemma,” in Proc. 25th Eur. Signal Process. Conf., 2017, pp. 698–702.

[12] Loic Bontemps, Van Loi Cao, James McDermott, and Nhien-An Le-Khac. Collective Anomaly Detection based on Long Short Term Memory Recurrent Neural Network.

[13] Ellis, Daniel PW. "Beat tracking by dynamic programming." Journal of New Music Research 36.1 (2007): 51-60. http://labrosa.ee.columbia.edu/projects/beattrack/

[14] Grosche, P., & Muller, M. (2011). "Extracting predominant local pulse information from music recordings." IEEE Transactions on Audio, Speech, and Language Processing, 19(6), 1688-1701.

[15] D. Yul Oh and I. Yun, “Residual error based anomaly detection using auto-encoder in smd machine sound,” Sensors, vol. 18, 201

[16] Preuveneers, D., Tsingenopoulos, I., & Joosen, W. (2020). Resource Usage and Performance Trade-offs for Machine Learning Models in Smart Environments. *Sensors (Basel, Switzerland)*, *20*(4), 1176. https://doi.org/10.3390/s20041176

[17] Jules Muhizi. Internal Publication, Fermi National Accelerator Laboratory.

[18] Banbury, Colby & Janapa Reddi, Vijay & Lam, Max & Fu, William & Fazel, Amin & Holleman, Jeremy & Huang, Xinyuan & Hurtado, Robert & Kanter, David & Lokhmotov, Anton & Patterson, David & Pau, Danilo & Seo, Jae-sun & Sieracki, Jeff & Thakker, Urmish & Verhelst, Marian & Yadav, Poonam. (2020). Benchmarking TinyML Systems: Challenges and Direction.