Anomalous Event Detection in Traffic Audio

Minakshee Shukla, Renu M Rameshan and Shikha Gupta
Vehant Research Lab, Vehant Technologies Pvt. Ltd., Noida, India

Keywords: Sound Event Detection, Monophonic, Teacher-Student Strategy, ATST, CRNN.

Abstract: This work focuses on detecting anomalous sound events from traffic audio. The audio used is recorded from the microphone associated with a surveillance camera. We have defined six anomaly classes and generated synthetic data using real background audio which corresponds to Indian traffic sound obtained from a surveillance camera microphone. Using a teacher-student training strategy, we have obtained F1 score of 96.12% and an error rate 0.06. We also show that even when the event occurs farther away from the microphone, the performance is still impressive, with an F1 score of 92.55 and an error rate of 0.12.

1 INTRODUCTION

Human auditory system has evolved over millions of years to have the ability to identify the different sounds that exist in the environment even when they are all mixed together. It understands when and from which direction the sound comes under most circumstances. But this is not true for a man made system. Systems that can understand the occurrence of a particular sound event in the presence of other signals are still scarce. In sound event detection one needs to detect the occurrence of a particular sound: when it occurred, when it ended, as well as what it is. This work focuses on detecting sound events which are anomalous with respect to a traffic scene, where there is a constant background of traffic noise which makes the task challenging.

Unlike speech and music, where there is a statistical structure, an audio event, more so an anomalous event, doesn't have any such structure. This makes the conventional approaches based on Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) or those with SVM classifiers, inefficient. The past few years has seen a prolific rise of deep learning based solutions for sound event detection (SED) as well as for anomalous/rare sound event detection.

In this preliminary study, we have chosen to work with six anomaly classes. Our classes of interest are car crash, tire skidding, glass breaking, gunshot, explosion, and screaming. Such a system can promptly alert emergency response teams in the event of a mishap to help those who are affected. This work also can be useful for acoustic monitoring of natural or artificial environments including but not limited to smart cities, for noise monitoring for safe and healthy living as well as to manage the noise pollution for safe hearing. Sound is not a conventional signal for anomaly detection in traffic surveillance which is usually done using videos. A video based surveillance system fails in the cases of fog, cloud, snow, smoke, occlusions etc. and in such cases an audio based system can take over the anomaly detection.

Considering the rare nature of the anomaly classes that we have considered, we design a monophonic sound event detection system. The main challenges in designing a learning based solution is the lack of strongly labelled data. Annotating real traffic data is costly and time intensive, leading to the issue of insufficient labeled data for the anomaly classes we selected for our use case. Also lack of realistic data for the classes of interest is another major issue. In order to solve these issues for the classes that we selected, we created a dataset for these six classes, using real traffic background sound and also used a collection of strongly labeled, weakly labeled and unlabeled samples.

A transformer based student-teacher network has been used for Sound Event Detection (SED) in (Shao et al., 2023). We adapt this method and retrain it for our purpose. We achieve a 99.33% accuracy. In addition, we also show, via simulation that the performance of the system does not degrade much with the distance of occurrence of the event from the camera.

Shukla, M., Rameshan, R. M. and Gupta, S.
Anomalous Event Detection in Traffic Audio.
DOI: 10.5220/0013152100003905
In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 613-619
ISBN: 978-989-758-730-6; ISSN: 2184-4313

2 RELATED WORKS

A brief overview of related work is given in this section. Prior research in audio processing has explored various supervised, unsupervised and self-supervised learning approaches. Supervised methods rely on labeled data, which can be expensive and time consuming to obtain. Unsupervised and self-supervised learning methods are gaining popularity due to their ability to learn from patterns or by generating representations for audio samples using contrastive loss (Khosla et al., 2020).

Methods which used traditional machine learning worked with Gaussian Mixture Models (GMM) (Ito et al., 2009), Hidden Markov Models (HMM) for feature extraction and classification (Zeiger, 2008) or support vector machines (Foggia et al., 2016) with features based on MFCC. (Giri et al., 2020) used pool of support vector machines (SVM) to detect two classes of hazardous road events tire skidding and car-crash.

As mentioned in the previous section, these traditional methods do not perform well when it comes to the anomalous sound event detection task.

Recent approaches rely on convolutional neural networks, specifically models like auto encoders, and variational auto encoders. Also, transformer based and knowledge distillation based solutions exist for sound event detection. Two of the auto encoder based solutions are (Koizumi et al., 2019b) and (Wichern et al., 2021). In the former an auto encoder was trained to minimize the reconstruction error of observed sounds, thus reducing the false positive rate. They used synthetic data. (Wichern et al., 2021) used an objective function based on Neyman Pearson lemma (Neyman and Pearson, 1933) to train the auto encoder to maximize the true positive rate. Wichern et al. used attentive neural processes, a meta learning approach, to train a masking based auto encoder for audio anomaly detection. They have reported the results on publicly available anomaly sound detection dataset for machine sounds, MIMII (Purohit et al., 2019). MIMII is an open-source dataset for malfunctioning industrial machine investigation and inspection.

Another class of approaches that we explored include large models specifically for feature generation. (Kong et al., 2020) designed PANNs (pretrained audio neural networks) by training a CNN model called wavegram-logmel-CNN using both log-mel spectrograms and waveform as input feature, though, it gives SOTA results for the audio tagging tasks for 527 audioset (Gemmeke et al., 2017) classes of sound. Similar performance is achieved with another open source model named YAMNet (YAM,). YAMNet used MobileNetV1 (Howard et al., 2017) architecture and was trained with audioset data (Gemmeke et al., 2017) for providing inference for 521 predefined classes of audioset. While both the models performed well for audio tagging, the performance on anomalous sound event detection was not good in the presence of background noise.

Though natural sounds occur in a polyphonic manner, because of the way we model our problem, we treat the sound samples as monophonic. The work in (Radford et al., 2023) is a supervised approach dealing with polyphonic sound event detection. This work proposes a convolutional recurrent neural network (CRNN) which combines the ability of CNNs to extract high level spatio-temporal invariant features and the ability of RNN to learn long term temporal correlations.

In recent years, DCASE (DCA,) has attracted many researchers to the domain of acoustic event and scene detection task which led to multiple novel solutions to this problem. Some of the recent ones being DCASE 2023 task 4 (DCA, 2023) baseline provided by the challenge organizers, which uses BEATs (Chen et al., 2023) (Bidirectional Encoder representation from Audio Transformers) and a CRNN network in a teacher-student manner for sound event detection task in domestic environment using DESED dataset (DES,). A paper on similar lines is, Shao et al. (Shao et al., 2023) which replaced BEATs in (Chen et al., 2023) with ATST-Frame model introduced in (Li et al., 2023). We are using this idea and model in our proposed anomalous sound event detection for traffic audio.

Another recent work is SPecific anomaly IDentifiER network called SPIDERnet (Koizumi et al., 2020) which is a one-shot anomaly detection method for anomalous sound. In this anomaly detection system they used a neural network-based feature extractor for measuring similarity in embedded space and attention mechanisms for absorbing time-frequency stretching. Although SPIDERnet outperforms conventional methods and robustly detects various anomalous sounds we decided not to adapt this solution as the results are provided only for the machine datasets ToyADMOS (Koizumi et al., 2019a) and MIMII (Purohit et al., 2019).

3 PROPOSED SOLUTION

The most general approach to anomaly detection is to train a system to learn what is normal and anything which falls out of the distribution of normal is marked

as anomalous. Anomalous events are usually rare, and unknown apriori, making the learning process quite challenging. In certain restricted cases where we know the types of anomalies that could come up, one can follow a classification based approach for anomaly detection. Our focus being anomaly detection in Indian traffic, we select six possible classes and train a system to raise an alarm when any of these six events occur. The classes considered are 1. car crash 2. tire skidding 3. gun shot 4. explosion 5. glass breaking and 6. screaming .

In this paper we have followed the work of Shao et al. (Shao et al., 2023) which finetunes the pre-trained ATST model for sound event detection task. We use the architecture from (Li et al., 2023) and apply the finetuning strategy of (Shao et al., 2023) for training using our dataset. Our focus is on monophonic multi-class traffic anomaly detection along with recognizing the start time i.e onset and end time i.e offset of the anomaly.

The architecture consists of the pretrained ATST-Frame (Li et al., 2023) model along with a CRNN network. ATST-Frame is an audio teacher-student transformer based model architecture which is trained in such a way that it can learn frame-wise representations by maximizing the similarity between student's and teacher's frame-level embeddings. It is important to note that, Shao et al. replaced pre-trained BEATs (Chen et al., 2023) with pre-trained ATST-frame model from the baseline solution of DCASE 2023 challenge (DCA, 2023) and the got SOTA results for sound event detection task using DESED dataset (DES,).

The idea of the paper (Shao et al., 2023) is to use a trained CRNN to finetune the ATST model. Training progresses in two stages: in stage 1 the ATST model is kept frozen and the CRNN is trained using the labelled samples via BCE loss and the output of ATST via the Mean Teacher (MT) loss. In the second stage of training the output of CRNN acts as pseudo label data for finetuning the ATST. In this step the training loss is a combination of binary cross entropy (BCE), mean-teacher (MT) and interpolation consistency training (ICT). It may be noted that both the models are updated in the latter stage.

The system has been trained using strongly labelled, weakly labelled and unlabelled data. The process of data creation is given in the next section.

4 DATASET CREATION AND TRAINING

We have prepared the dataset for sound event detection problem with explicitly collected background audios from specified traffic sites and event only audios collected from various open-source datasets including SESA (Sound Events for Surveillance applications) (SES,) MIVIA labs audio events dataset (MIV, a), Mivia Labs ROAD Audio Events Dataset (MIV, b) and DCASE 2017, TUT Rare Sound Events 2017 (DCA, 2017).

4.1 Real Background Audio Data Creation

The background audios are recorded at a single site. Nevertheless, a natural variability is introduced by the varying congestion conditions that exists across the day. In addition, we have introduced two different SNR levels: 0dB and 5dB for the background audio.

For recording the background sound, we used in-built microphone from an IP camera device DH-IPC-HF5231EP-E 2MP WDR Box Network Camera (Dah,) from Dahua Security. This camera is installed at a traffic junction in a busy Indian city. Background data is recorded in .AAC format at 16KHz sampling rate and saved on an hourly basis. One hour long background data is then converted from .aac to .wav format and splitted into segments of duration 10 second each.

4.2 Event Audio Collection

The anomalous event audios were extracted from their respective datasets and the event alone part was clipped out based on the onset and offset information given in their respective meta data. These are, then down-sampled to 16kHz audio segments.

4.3 Background and Event Mixing Procedure

After the above mentioned steps, the resulting data had a total of 22,225 background audio samples of 10 seconds each and a total of 670 event audio files. The six event classes are Carcrash, Gunshot, Glassbreaking, Screaming, Explosion and Tireskidding. These audio events were then synthetically mixed with the traffic environment background audios.

Based on the energy level of the audio clips the background data is split into three categories of low, medium and high energy audio files. These three

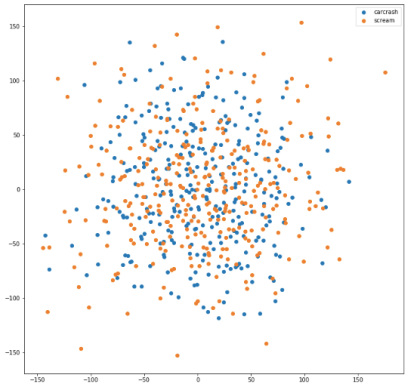

Figure 1: t-SNE scatter plot for Carcrash and Screaming using PANNs embeddings.

Table 1: Segment Based Performance Metrics.

Class	Stage 1		Stage 2	
	Precision	Recall	Precision	Recall
Tireskidding	87.6	77.3	97.0	94.8
Carcrash	59.0	78.5	96.7	92.9
Screaming	52.3	91.9	96.7	98.9
Explosion	98.7	97.7	98.2	98.9
Glassbreaking	74.0	66.5	79.8	98.9
Gunshot	97.6	59.4	97.1	92.5
Overall	87.73	81.78	95.73	96.51
Overall Accuracy	97.45		99.33	
Overall F-score	84.65		96.12	
Overall Error-rate	0.22		0.06	

Table 2: Event Based Performance Metrics.

Class	Stage 1		Stage 2	
	Precision	Recall	Precision	Recall
Tireskidding	61.7	64.0	83.3	89.5
Carcrash	31.7	55.8	90.9	87.9
Screaming	41.2	76.5	100	100
Explosion	94.0	95.5	97.4	98.0
Glassbreaking	51.0	51.7	70.2	96.5
Gunshot	75.2	37.4	98.5	90.8
Overall	66.11	61.74	92.74	93.86
Overall F-score	63.85		93.30	
Overall Error-rate	0.62		0.10	

Table 3: Segment Based Performance Metrics for Stage 2 (SNR 0 dB).

Class	Event Scale 1		Event Scale 0.25	
	Precision	Recall	Precision	Recall
Tireskidding	97.1	94.4	97.2	90.6
Carcrash	96.8	91.9	95.3	78.7
Screaming	96.5	98.9	97.8	97.8
Explosion	98.3	98.7	97.2	93.8
Glassbreaking	78.9	99.5	77.7	89.1
Gunshot	96.8	92.4	97.0	85.7
Overall	95.53	96.42	94.99	90.23
Overall Accuracy	99.30		98.74	
Overall F-score	95.97		92.55	
Overall Error-rate	0.06		0.12	

background audio categories and each of the class-wise event data are split in the ratio of 6:2:2 for train set, validation set and test set, respectively. This energy based background audio splitting ensured that all three levels of Indian traffic congestion conditions - low, medium and highly noisy - are equally distributed across the training, validation and testing sets of the data created.

Table 4: Event Based Performance Metrics for Stage 2 (SNR 0 dB).

Class	Event Scale 1		Event Scale 0.25	
	Precision	Recall	Precision	Recall
Tireskidding	83.7	88.1	87.4	87.1
Carcrash	91.6	87.5	75.9	70.9
Screaming	100	100	96.8	98.0
Explosion	96.6	97.6	60.7	76.1
Glassbreaking	68.7	95.6	63.0	84.4
Gunshot	98.8	90.8	88.3	82.0
Overall	92.48	93.49	75.59	81.26
Overall F-score	92.98		78.32	
Overall Error-rate	0.11		0.41	

Table 5: Performance Metrics for CRNN only Network.

Class	Segment Based		Event Based	
	Precision	Recall	Precision	Recall
Tireskidding	75.0	98.1	71.8	85.7
Carcrash	74.1	75.6	29.2	27.2
Screaming	81.6	91.7	69.9	82.7
Explosion	91.0	99.1	71.3	74.8
Glassbreaking	61.6	80.5	40.1	40.1
Gunshot	58.6	90.0	62.1	67.5
Overall	74.97	93.08	61.86	66.19
Overall F-score	83.05		63.95	
Overall Error-rate	0.35		0.71	

The event only signal duration varies from 0.5s up to 9.9s with varying duration for different classes. Especially, event segments like tire skidding have comparatively longer duration. As the background audio duration is 10s, total duration of the mixture is fixed to 10s. Mixing of the background and event are done independently in all the three categories, thus ensuring that no sample is common across the three sets: train, validation and test.

The background and the event are mixed at two SNR levels 0dB and 5dB. Let $e[n]$ be the event and $b[n]$ be the background, the scaling factor a is calculated as:

$$a = \sqrt{\frac{\sum_{i=0}^{N-1} e^2[i]}{10^{\frac{(SNR)}{10}} \sum_{i=0}^{M-1} b^2[i]}} \quad (1)$$

As N, M, the total length of event and background, respectively, are different with the condition $M > N$, we need to decide, where in the duration 0 to $M-N$, we need to add the event. For this, we select an event audio file and a corresponding background audio file into which the event file is added at a random location between 0 to $M-N$.

We mixed event and background data in such a way that we could prepare the data according to these

segments: strongly labelled data, weakly labelled data and unlabelled data.

In strongly labelled data, for each mixed audio file event the start time (onset) and event end time (offset) and event label (class to which it belongs) is given. In weak labels, only the event class label is available, and in unlabelled segments we have only mixed files with no other information, and these form the self-supervised data.

Total of 34716 files have been created, out of which 21256 are used for training, 9001 for testing and 4459 audio files for validation.

4.4 Training

Training was done in two different stages because both the models required very different training schedule and learning parameters.

In stage 1, ATST-Frame was kept frozen and only CRNN network was trained with batch size 4,8,8 for strongly labelled, weakly labelled and unlabelled data. The input audio clips were divided in frames of duration 128*ms* with a hop length 16*ms*. 128-dimensional Log-Mel features were extracted for each frame.

Data augmentation is randomized with no augmentation having a probability of 0.5 with mix-up and frequency-warping, each having a probability of 0.25, each. For stage 1, learning rate was set at 1e-3 for both CNN and RNN.

Same batch size was used for stage 2 as well. For stage 2, learning rate was set at 2e-4 and 2e-3 for CNN and RNN, respectively.

Adam optimizer has been used. Binary cross entropy loss and both mean teacher loss and interpolation consistency training loss has been used for supervised and unsupervised loss calculation, respectively.

5 RESULTS

We evaluated the above system on a total of 9001 audio clips. The evaluation has been performed using two methods: event based and segment based, inspired from polyphonic sound detection score metric calculation (Bilen et al., 2020), where event wise evaluation provides clip level evaluations, segment wise results provide evaluation score based on frame length of 64 ms. For segment based results, we have calculated the overall accuracy (calculated based on micro average) for the system which is going from 97.45 percent to 99.33 percent in stage 1 and stage 2, respectively.

Table 1 gives segment-wise precision and recall for stage 1 and stage 2. It may be noted that both precision and recall have increased considerably for all the 6 classes after stage 2 training. Table 2 gives the event based precision and recall for both the stages. Here also a remarkable improvement in the performance is seen after stage 2 training.

In order to see the performance when the location, at which the event occurred is far away from the mic, we created a test data set where the event amplitudes were scaled by a factor of 0.25. To compare the performance between original and the scaled audio, we created two test datasets, one with event scale 1 and the other with event scale 0.25. Everything else, including the onset and offset of the events is same between the two event scales for fair comparison. Table 3 gives the stage 2 results for segment-based evaluation, when the event source moves away. It is seen that, precision falls only slightly, indicating that the true positives are still predicted correctly to a large extent. Whereas the fall in recall, indicates that many of the true cases are missed, being classified as false negatives. This is also reflected in the error score.

Similarly, Table 4 gives the stage 2 results for event-based evaluations for event scale 0.25. Here, the performance degrades, significantly. Both the false positives and false negatives have increased considerably as is evident from the lower precision and recall as compared to the corresponding event scale 1 values. The error has risen to 0.41 in line with there observations.

It may be noted that all the overall metric scores mentioned in the tables (1-5) have been calculated on the basis of micro-averaging.

To establish that the student-teacher based model is far superior, we also show the results for a simple CRNN in Table 5. Comparison with other models are not made available as the embeddings learnt using those (PANNs, YAMNet) were not discriminative enough. It was observed from the t-SNE plot that these anomalous classes overlapped with each other. In Fig. 1 we show the tSNE of PANNs (Kong et al., 2020) embeddings for two classes: screaming and carcrash. It is evident from the plot that classification is nearly impossible. From table 5, it is seen that, segment-based as well as event-based scores for fine-tuned ATST-SED model on traffic anomaly classes are far superior to the base CRNN results.

6 CONCLUSIONS

In this paper, we re-trained a transformer based student-teacher network for Sound Event Detection

task following Shao et al. (Shao et al., 2023), using synthesized dataset for our purpose. Based on our model, unknown acoustic patterns are identified into six different anomaly classes. The use of student-teacher transformer allows the learning of long-term temporal dependencies. When trained and fine-tuned with on the synthetic dataset generated using real traffic audio, the model gave an overall accuracy of 99.33% when tested on unseen audio. The model performance degrades gracefully with distance of the source of anomaly which is an added advantage.

REFERENCES

Dahua security network cameras. https://zenodo.org/records/3519845. Accessed: October 22, 2024.

Dahua security network cameras. https://www.dahuasecurity.com/in/products/All-Products/Network-Cameras/WizMind-Series/5-Series/2MP/DH-IPC-HF5231EP-E. Accessed: October 22, 2024.

DCASE community. https://dcase.community/. Accessed: October 22, 2024.

DESED: Domestic environment sound event detection. https://project.inria.fr/desed/. Accessed: October 22, 2024.

MIVIA audio events dataset. https://mivia.unisa.it/datasets/audio-analysis/mivia-audio-events. Accessed: October 22, 2024.

MIVIA road audio events dataset. https://mivia.unisa.it/datasets/audio-analysis/mivia-road-audio-events-data-set. Accessed: October 22, 2024.

YAMNet. https://www.tensorflow.org/hub/tutorials/yamnet. Accessed: October 22, 2024.

(2017). Dcase 2017 challenge: Rare sound event detection. https://dcase.community/challenge2017/task-rare-sound-event-detection. Accessed: October 22, 2024.

(2023). DCASE 2023 challenge: Sound event detection with weak and soft labels. https://dcase.community/challenge2023/task-sound-event-detection-with-weak-and-soft-labels. Accessed: October 22, 2024.

Bilen, C., Ferroni, G., Tuveri, F., Azcarreta, J., and Krstulovic, S. (2020). A framework for the robust evaluation of sound event detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 61–65.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X., and Wei, F. (2023). Beats: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning (ICML)*, pages 5178–5193.

Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., and Vento, M. (2016). Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE Transactions on Intelligent Transportation Systems*, 17(1):279–288.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Giri, R., Tenneti, S. V., Cheng, F., Helwani, K., Isik, U., and Krishnaswamy, A. (2020). Self-supervised classification for detecting anomalous sounds. In *Proc. DCASE*, pages 46–50.

Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Ito, A., Aiba, A., Ito, M., and Makino, S. (2009). Detection of abnormal sound using multi-stage gmm for surveillance microphone. In *Proc. IAS*, volume 1, pages 733–736.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. In *Proceedings of the Advances in Neural Information Processing Systems*.

Koizumi, Y., Saito, S., Uematsu, H., Harada, N., and Imoto, K. (2019a). Toyadmos: A dataset of miniature-machine operating sounds for anomalous sound detection. In *Proc. of the Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

Koizumi, Y., Saito, S., Uematsu, H., Kawachi, Y., and Harada, N. (2019b). Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):212–224.

Koizumi, Y., Yasuda, M., Murata, S., Saito, S., Uematsu, H., and Harada, N. (2020). Spidernet: Attention network for one-shot anomaly detection in sounds. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 281–285, Barcelona, Spain.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (2020). Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *arXiv:1912.10211*.

Li, X., Shao, N., and Li, X. (2023). Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks. *arXiv preprint arXiv:2306.04186*.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London*, 231:289–337.

Purohit, H., Tanabe, R., Ichige, K., Endo, T., Nikaido, Y., Suefusa, K., and Kawaguchi, Y. (2019). Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, page 209.

Radford, A., Narasimhan, K., Jeong, D., Chowdhery, J., Butler, J., Shuster, L., Parmar, N., Clark, D., and Elibol, J. (2023). Robust speech encoding via large-scale weak supervision. *arXiv preprint arXiv:2309.06864v2*.

Shao, N., Li, X., and Li, X. (2023). Fine-tune the pretrained atst model for sound event detection. *arXiv preprint arXiv:2309.08153v2*.

Wichern, G., Chakrabarty, A., Wang, Z. Q., and Roux, J. L. (2021). Anomalous sound detection using attentive neural processes. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 186–190, New Paltz, NY, USA.

Zeiger, C. (2008). An HMM based system for acoustic event detection. In *Multimodal Technologies for Perception of Humans*.