# Review of anomalous sound event detection approaches

**Amirul Sadikin Md Afendi[1], Marina Yusoff[2]**
[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
[2]Advanced Analytic Engineering Center (AAEC), Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

**Article Info**

***Article history:***

Received Jun 27, 2019
Revised Aug 5, 2019
Accepted Aug 20, 2019

***Keywords:***

Acoustic Features
Audio Processing
Pattern Recognition
Sound Event Detection

## ABSTRACT

This paper presents a review of anomalous sound event detection (SED) approaches. SED is becoming more applicable for real-world appliactaions such as security, fire determination or olther emergency alarms. Despite many research outcome previously, further research is required to reduce false positives and improve accurracy. SED approaches are comprehensively organized by methods covering system pipeline components of acoustic descriptors, classification engine, and decision finalization method. The review compares multiple approaches that is applied on a specific dataset. Security relies on anomalous events in order to prevent it one must find these anomalous events. Audio surveillance has become more efficient as that artificial intelligence has stepped up the game. Autonomous SED could be used for early detection and prevention. It is found that the state of the art method viable used in SED using features of log-mel energies in convolutional recurrent neural network (CRNN) with long short term memory (LSTM) with a verification step of thresholding has obtained 93.1% F1 score and 0.1307 ER. It is found that feature extraction of log mel energies are highly reliable method showing promising results on multiple experiments.

***Corresponding Author:***

Marina Yusoff,
Advanced Analytic Engineering Center (AAEC),
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia.
Email: marinay@tmsk.uitm.edu.my

## 1. INTRODUCTION

The Sound event detection (SED) research field has been an active recently [1-4]. Autonomous audio surveillance has become more efficient as that artificial intelligence has stepped up the game [5]. Machine learning has become very powerful to allow new approaches to be facilitated in countless domain [6]. Acoustic surveillance specifically in security application is still new to the world and requires research to allow better performances [4, 7, 8]. The detection on anomalous audio events effectively while avoiding false positives is crucial in security [3]. Two categories of sound recognition, non-speech the determination of the sound event source and speech recognition of verbal language [9-10]. Many works embrace the method of multi-label classification to detect polyphonic acoustic events with a worldwide limit to detect active acoustic events [2, 7, 10]. The lack of accuracy and high false positives on current SED approaches is holding autonomous audio surveillance to be applied for real-world applications [11]. In the race towards making audio surveillance a reality many have aid by providing data for training and testing new models in solving the problem. Datasets on SED can be found available in many varieties as the internet of things are becoming more popular [1, 10, 12].

The rare term relates to the possibility that any target sound event to be identified may happen at most once within half a minute [4, 12]. Detecting rare sound events are one of the key features for security [10]. The approaches used in previous studies, particularly in detection of rare events can be simplified by the combinations of certain parts of the whole detection system [8]. Parts of the detection system that consists of the audio channel input data, feature extraction, classification, and decision making [8]. This paper presents a review of anomalous sound event detection approaches by organizing a number of research done on anomalous SED recently. Comparing them on the basis of the whole system composition towards the results achieved. Observations are mainly on the composition of the typical SED pipeline components. The performance of the system are evaluated on a single and multi-class basis.

## 2. SOUND EVENT DETECTION

The job of identifying sound events includes identifying and classifying sounds in audio forecasting and offset sounds for different cases of sound events and offering a textual descriptor for each [13-14]. Common classification method used on SED are as the conventional speech recognition [12]. Along the pipeline of a typical SED it consist of basically feature extraction followed by the classification [15]. Classification method that is popular recently consist of Convolution Neural Network (CNN) [16]. In surveillance SED architectures consist of additional background subtraction,object tracking & situational analysis processes in the pipeline [8]. Latest research of an improved pipeline suggest a verification step to reduce false positives after the SED process [3]. Figure 1 shows the extended SED pipeline.

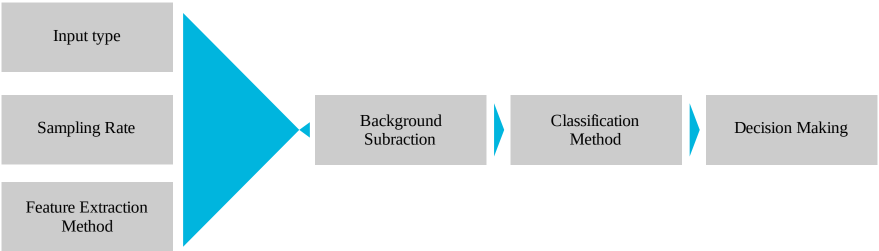

Figure 1. Sound event detection pipeline

In the first stage input type, sampling rate and feature extraction is the audio specifications of preprocessing step. The input type and sampling rate specification are grouped with feature extraction method because they are co-related, input type of mono or stereo channel and sampling rate will result of different amount of features extracted. Hence, it is combined as it is one process in the pipeline where proper tuning could be done. The audio is processed to typically 20-200 ms brief frames to obtain the audio characteristics of choice [10]. Smaller frames will make up more features could lead to overfitting in artificial intelligence systems, which can result in lower precision during classification [17]. Representations of the signal range, including such Mel-frequency coefficients (MFCC), Mel energies, or just the amplitude or energy range are often used in sound event identification [13-14].

### 2.1. Acoustic features

A major component in the SED pipeline is the feature extraction method. There are many features can be extracted from an audio source separated into two main types of feature spectral features and rhythm features. Common spectral features are as listed below.

1. Compute a chromagram from a waveform or power spectrogram.
2. Constant-Q chromagram
3. Mel-scaled spectrogram.
4. Mel-frequency cepstral coefficients (MFCCs)
5. root-mean-square (RMS) value
6. Compute the spectral centroid.
7. P’th-order spectral bandwidth.
8. Spectral contrast
9. Spectral flatness
10. Roll-off frequency.
11. coefficients of fitting an nth-order polynomial to the columns of a spectrogram
12. Compute the zero-crossing rate of an audio time series.

As cepstral characteristics are calculated by taking the transformation of the distorted logarithmic spectrum from Fourier, they contain data on the rate changes in the distinct spectrum bands. Because of their capacity to separate the effect of source and filter in a voice signal, Cepstral characteristics are beneficial for SED. Features could be employed best on their specifically designed purposes. Thus selection of features plays a vital role on SED systems [9]. Alternatives for tailored made feature extraction would be the raw waveforms acoustic modelling [18].

### 2.2. Featured dataset

The dataset featured in this review are based on the Detection Classification Acoustic Scene and Events (DCASE) 2017 Tampere Universiti Tecnology (TUT) Rare Sound Events 2017 development dataset [6]. The results of approaches compared are based on this dataset. Composition of audio materials sound events which are known are of the following classes:

1. Baby crying (106 instances for training, 42 instances for test)
2. Glass breaking (96 instances for training, 43 instances for test)
3. Gun shot (134 instances for training, 53 instances for test)

The statistical data of the sound event development datasets is as shown in Table 1.

Table 1. Featured dataset statistics

| Class | Baby Cry | | Glass Break | | Gun Shot | |
|---|---|---|---|---|---|---|
| Usage | Train | Test | Train | Test | Train | Test |
| Mean | 2.41 | 1.85 | 1.36 | 0.72 | 1.43 | 1.04 |
| Max | 5.1 | 4.24 | 4.54 | 0.182 | 4.4 | 3.68 |
| Min | 0.66 | 0.78 | 0.26 | 0.3 | 0.24 | 0.3 |
| Median | 2.33 | 1.67 | 1.29 | 0.7 | 1.21 | 0.76 |
| Standard Deviation | 0.98 | 0.83 | 0.75 | 0.3 | 0.86 | 0.83 |

The length of the target scenario is generally brief relative to the 30-second background noise [13]. Training and testing data sets are produced respectively. Combinations are produced by software package with source information, including ambient noise and specified occurrences [12].

### 2.3. Sed evaluation metric

The official metric in the DCASE2017 Challenge that is ER is the sum of insertion, deletion and substitution rates [7, 10]. ER is explained thoroughly in [10]. The evaluation involves calculation of true positive (TP), false positive (FP), and false negatives (FN). If the output of the system correctly predicts the presence and onset of a case, it will be calculated as TP. Only when it is predicted within 500 milliseconds (ms) of the actual starting time is detection as true [19]. FP demonstrates that if there is no event, the system wrongly detects the existence of an event. If the output of the strategy misses the event, an FN will be considered. These metrics are used in the final phase to calculate the error rate and the F-score.

$$ER = \frac{FN+FP}{N} \tag{1}$$

$$F = \frac{2PR}{P+R} \tag{2}$$

Where N indicates the total amount of samples in the assessment dataset and where P and R denote accuracy and recall as described below.

$$ER = \frac{FN+FP}{N} \tag{3}$$

$$F = \frac{2PR}{P+R} \tag{4}$$

## 3. PAST APPROACHES RESULTS

Traditionally SED popular approaches were Gaussian Mixture Model(GMM) and Hidden Markov Model(HMM) but has moved to the new deep learning-based methods [7, 20]. Recent approach uses deep learning-based methods that are popularly used consist of Deep Neural Network (DNN), Recurrent Neural

Network (RNN), Convolutional Recurrent Neural Network (CRNN) [21]. Table 2 shows past approaches based on a collection of the experiments conducted on the TUT Rare Sound Event 2017 Dataset is compared based on their respective accuracy, features, classifier and decision making specifications. The column features are the choice of features extraction method that will be fed into the classifier engine. The Classifier column denotes the method of detection employed in the system. The column decision making refers to the final verification step in the SED Pipeline. Additionally, F1 multiclass inconsistency show the deviation of each class performance. The approach best was the combination of a 1 Dimensional CNN and RNN with long short term memory units (LSTM) with the average accuracy of 93.1% [19].

Table 2. Results of past approaches of SED conducted on TUT Rare SED 2017

| Researchers | Overall | | F1 Multi-Class Inconsistency | Features | Classifier | Decision Making |
|---|---|---|---|---|---|---|
| | ER | F1 | | | | |
| [19] | 0.1307 | 93.1 | 4.1 | log-mel energies | CRNN-LSTM | thresholding |
| [7] | 0.1733 | 91 | 3.7 | spectrogram | CNN | majority vote |
| [3] | 0.2773 | 85.3 | 3.6 | log-mel energies | CRNN | median filtering ensemble |
| [24] | 0.3133 | 84.2 | 10.6 | log-mel energies | MLPCNN | theshold |
| [22] | 0.3173 | 82 | 0.9 | log Gammatone cepstral coefficients | tailored-loss DNN+CNN | median filtering |
| [21] | 0.3267 | 83.9 | 13.0 | log-mel energies | CRNN | majority vote |
| [11] | 0.4107 | 79.1 | 10.2 | log-mel SpectrogramsMFCC | MLPCNNRNN | median filtering ensembling hard Thresholding |
| [25] | 0.4267 | 78.6 | 8.3 | MFCCZCRenergyspectral centroidpitch | ensemble | thresholding |
| [12] | 0.432 | 73.4 | 20.0 | log-mel energies | DNN | median filtering |
| [5] | 0.5 | 74.2 | 17.9 | spectrogram | NMF | moving average filter |
| [7] | 0.6 | 69.8 | 11.6 | DNN(MFCC) | Bi-LSTM | top output probability |
| [6] | 0.6373 | 64.1 | 14.2 | log-mel energies | MLP | median filtering |
| [23] | 0.6773 | 65.8 | 16.5 | log-mel energies from NMF source separation | MLP | median filtering |

Furthermore we can observe the single class accuracy deviation below in Table 3. As you can see the differences between each class performance shows imbalance that could be related to the approach as an advantage. The deviation of an approach will show its single class performance could be employed to specifically targeted event. Table 3 shows the multi-class inconsistency in which is the deviation of the three target classes done in previous experiments.

Table 3. Multi-class inconsistency

| Researchers | Baby cry | | Glass break | | Gunshot | | F1 Multi-Class Inconsistency |
|---|---|---|---|---|---|---|---|
| | ER | F1 | ER | F1 | ER | F1 | |
| [19] | 0.152 | 92.2 | 0.048 | 97.6 | 0.192 | 89.6 | 4.1 |
| [7] | 0.184 | 90.8 | 0.104 | 94.7 | 0.232 | 87.4 | 3.7 |
| [3] | 0.284 | 85.7 | 0.22 | 88.8 | 0.328 | 81.6 | 3.6 |
| [22] | 0.172 | 91.4 | 0.22 | 89.1 | 0.548 | 72 | 10.6 |
| [11] | 0.356 | 83 | 0.312 | 84.7 | 0.312 | 84 | 0.9 |
| [21] | 0.264 | 87.3 | 0.16 | 91.5 | 0.528 | 67.2 | 13.0 |
| [25] | 0.44 | 80.6 | 0.228 | 88.5 | 0.564 | 68.2 | 10.2 |
| [24] | 0.5 | 75.9 | 0.236 | 87.8 | 0.544 | 71.9 | 8.3 |
| [5] | 0.408 | 78.8 | 0.164 | 91.5 | 0.928 | 52.3 | 20.0 |
| [12] | 0.44 | 77.3 | 0.212 | 89.1 | 0.644 | 53.9 | 17.9 |
| [2] | 0.78 | 67.4 | 0.324 | 82.4 | 0.696 | 59.5 | 11.6 |
| [23] | 0.884 | 65.3 | 0.396 | 80.2 | 0.752 | 51.8 | 14.2 |
| [6] | 0.804 | 66.8 | 0.38 | 79.1 | 0.728 | 46.5 | 16.5 |

Comparison of the experiment results that was done on a leveled dataset. The deviation of some approaches suggest the methods could be employed as a single class efficient. The combination of spectrogram, NMF and moving average filter [5] shows the highest deviation among each class performance. High deviation may indicate a specific class or pattern over fitting. The review believes that lower deviation should conclude a robust approach towards large class variations detection in the future. Large variety is the problem humans are able to compute. This deviation could be further researched to fully understand the trend and reasons behind such a phenomenon. Comparison between SED components combination suggest popular specifications for future research and development.

## 4. CONCLUSION

The study compares previous work on anomalous SED solutions observed mainly on the composition of the typical SED pipeline components. The performance of the system were evaluated on a single and multi-class basis. It is found that the performance on single class tend to be imbalanced and vary between approaches. The low deviation in single class performance should indicate a versatile approach that is not biased on a certain type of event. Thus the approach is believed to be well optimized for a wider variety of SED problems. In search for a robust SED pipeline components composition optimized for the best. It is suggested for further studies to employ the combination of system characteristics of such feature extraction, classifiers and decision making steps in the pipeline. The most viable approach in this review is that a combination of using features as log-mel energies in convolutional recurrent neural network(CRNN) with long short term memory(LSTM) with a verification step of thresholding has obtained 93.1% F1 score and 0.1307 ER. Future work is suggested to be revolving around these specifications making up the entire pipeline to produce a reliable product for real-world application.

## ACKNOWLEDGEMENTS

Universiti Teknologi MARA a for the grant of 600-IRMI/PERDANA 5/3 BESTARI (096/2018) as well as Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia for providing essential support and knowledge for the work.

## REFERENCES

[1] Boddapati, V., Petef, A., Rasmusson, J., & Lundberg, L. (2017). Classifying environmental sounds using image recognition networks. *Procedia Computer Science*, 112, 2048–2056. https://doi.org/10.1016/j.procs.2017.08.250.

[2] Li, Y., & Li, X. (2017). The {SEIE-SCUT} Systems for {IEEE} {AASP} Challenge on {DCASE} 2017: Deep Learning Techniques for Audio Representation and Classification.

[3] Phan, H., Koch, P., Katzberg, F., Maass, M., Mazur, R., McLoughlin, I., & Mertins, A. (2017). What makes audio event detection harder than classification? 25th European Signal Processing Conference, EUSIPCO 2017, 2017-January, 2739–2743. https://doi.org/10.23919/EUSIPCO.2017.8081709

[4] Yusoff, M & M. Afendi A. S. (2019). Acoustic Surveillance Intrusion Detection with Linear Predictive Coding and Random Forest: 4th International Conference, SCDS 2018, Bangkok, Thailand, August 15-16, 2018, Proceedings. 10.1007/978-981-13-3441-2_6.

[5] Ghaffarzadegan, S., Salekin, A., Das, S., & Feng, Z. (2017). Bosch Rare Sound Events Detection Systems for {DCASE2017} Challenge.

[6] Heittola, T., & Mesaros, A. (2017). {DCASE} 2017 Challenge Setup: Tasks, Datasets and Baseline System.

[7] Cakir, E., & Virtanen, T. (2017). Convolutional Recurrent Neural Networks for Rare Sound Event Detection.

[8] Crocco, M., Cristani, M., Trucco, A., & Murino, V. (2014). Audio Surveillance: a Systematic Review, (October). Retrieved from http://arxiv.org/abs/1409.7787

[9] Ozer, I., Ozer, Z., & Findik, O. (2018). Noise robust sound event classification with convolutional neural network. *Neurocomputing*, 272, 505–512. https://doi.org/10.1016/j.neucom.2017.07.021

[10] Jayalakshmi, S. L., Chandrakala, S., & Nedunchelian, R. (2018). Global statistical features-based approach for Acoustic Event Detection. Applied Acoustics, 139(April), 113–118. https://doi.org/10.1016/j.apacoust.2018.04.026

[11] Vesperini, F., Droghini, D., Ferretti, D., Principi, E., Gabrielli, L., Squartini, S., & Piazza, F. (2017). A Hierarchic Multi-Scaled Approach for Rare Sound Event Detection.

[12] Wang, Jun, & Li, S. (2017). Multi-Frame Concatenation for Detection of Rare Sound Events Based on Deep Neural Network.

[13] Mesaros, A., Heittola, T., & Virtanen, T. (2016). Metrics for Polyphonic Sound Event Detection. Applied Sciences, 6(6), 162. https://doi.org/10.3390/app6060162

[14] Bezoui, M. (2019). Speech Recognition of Moroccan Dialect Using Hidden Markov Models. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 8(1), 7. https://doi.org/10.11591/ijai.v8.i1.pp7-13

[15] Wang, Jianfei, Zhang, W., & Liu, J. (2017). Transfer Learning Based {DNN}-{HMM} Hybrid System for Rare Sound Event Detection.

[16] Phan, H., Krawczyk-Becker, M., Gerkmann, T., & Mertins, A. (2017). {DNN} and {CNN} with Weighted and Multi-Task Loss Functions for Audio Event Detection.

[17] Kamala, R., & Thangaiah, R. J. (2019). An Improved Hybrid Feature Selection Method for Huge Dimensional Datasets. *IAES International Journal of Artificial Intelligence (IJ-AI),* 8(1), 77. https://doi.org/10.11591/ijai.v8.i1.pp77-86

[18] Y. Hoshen, R. J. Weiss and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, 2015, pp. 4624-4628. https://doi: 10.1109/ICASSP.2015.7178847

[19] Lim, H., Park, J., & Han, Y. (2017). Rare Sound Event Detection Using {1D} Convolutional Recurrent Neural Networks.

[20] Sak, H., Senior, A., Rao, K., & Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2015-Janua, 1468-1472.
[21] Kaiwu, W., Liping, Y., & Bin, Y. (2017). Audio Events Detection and Classification Using Extended {R-FCN} Approach.
[22] Zhou, Q., & Feng, Z. (2017). Robust Sound Event Detection through Noise Estimation and Source Separation Using {NMF}.
[23] Jeon, K. M., & Kim, H. K. (2017). Nonnegative Matrix Factorization-Based Source Separation with Online Noise Learning for Detection of Rare Sound Events.
[24] Ravichandran, A., & Das, S. (2017). Bosch Rare Sound Events Detection Systems for {DCASE2017} Challenge.
[25] Dang, A., Vu, T., & Wang, J.-C. (2017). Deep Learning for {DCASE2017} Challenge.

## BIOGRAPHIES OF AUTHORS

Muhamad Amirul Sadikin MD Afendi is a Master of Science (Information Technology) student in Universiti Technologi MARA (UiTM) Shah Alam, Malaysia. Since 2018 has begun research in the field of artificial intelligence specifically on audio processing. His education background consist of a Bachelor of Information technology (Hons) Intelligent System Engineering at UiTM Shah Alam,Malaysia bythe year 2019. He is currently pursuing research on audio based forest surveillance collaborating with Wildlife Conservative Society Malaysia to combat wildlife abuse.

Marina Yusoff is currently a Deputy Dean (Research and Industry Linkages) and Senior Lecturer of Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Malaysia. She holds a PhD in Intelligent System from the Universiti Teknologi MARA. She previously worked as a senior executive of Information Technology in SIRIM Berhad, Malaysia. She holds a bachelor’s degree in computer science from the University of Science Malaysia, and Master of Science in Information Technology from Universiti Teknologi MARA. She is interested in the development of intelligent application, modification and enhancement computational intelligence techniques include particle swarm optimisation, neural network, genetic algorithm, and ant colony. She has many impact journals publications and presented her research in many conferences locally and internationally.