Audio event detection using deep learning model

E. Sophiya*

Department of Computer Science and Engineering,
Faculty of Engineering and Technology (FEAT),
Annamalai University,
Annamalainagar, India
Email: venus.sophiya@gmail.com
*Corresponding author

S. Jothilakshmi

Department of Information Technology,
Faculty of Engineering and Technology (FEAT),
Annamalai University,
Annamalainagar, India
Email: jothi.sekar@gmail.com

Abstract: Humans are surrounded by a complex audio stream that carries meaningful information about our surroundings in day to day life. Hearing is one of the most important capabilities to identify and detect audio events in our surroundings. For example sounds such as ambulance siren, gunshot, baby cry, etc which require immediate action. Thus, automatic audio analysis is getting popular in recent years which have wide range of applications such as continuous monitoring for public safety, abnormal events, wildlife monitoring, healthcare, audio indexing and retrieval. The objective of the proposed system is to provide the event class and the event time boundaries between multiple events present in an audio. The proposed audio event detection is implemented with a deep learning model. The real time data are collected from major locations of an urban city. Audio events were recognised using signal processing techniques. The model is learned from Log Mel spectrogram features.

Keywords: audio processing; audio scene analysis; audio event detection; deep learning; deep convolutional neural network.

Reference to this paper should be made as follows: Sophiya, E. and Jothilakshmi, S. (2022) 'Audio event detection using deep learning model', *Int. J. Computer Aided Engineering and Technology*, Vol. 16, No. 3, pp.328–343.

Biographical notes: E. Sophiya earned her Bachelor's degree in Computer Science and Engineering from Ponnaiyah Ramajayam College of Engineering and Technology. She received her Master's degree in Computer Science and Engineering from Annamalai University. She has working experience as an Assistant Professor in Christ College of Engineering and Technology, Puducherry. She is currently pursuing PhD (full time) in Computer Science and Engineering at Annamalai University. Her research is focused on audio scene classification, event detection, large scale data, machine learning, and deep learning.

S. Jothilakshmi earned her Bachelor's degree in Electronics and Communication Engineering from Madras University. She received her Doctoral and Master's degrees in Computer Science and Engineering from Annamalai University. She was a postdoctoral researcher at Marshall University, USA. She currently works as Associate Professor in the Department of Information Technology at Annamalai University. She has 19 years of teaching experience. She worked with her nationally funded research project and published more than 30 research articles in speech and image processing, machine learning, information retrieval, and big data. She authored a book titled *Communication Engineering: Theory and Concepts*. Three PhD students completed their research under her guidance. Currently, five students are pursuing their research.

1 Introduction

Sound is one of the most important media for communication which highly attracts the attention. Humans have the capability to recognise the environment into meaningful categories. Clustering the environments of similar categories will help to handle the complex environments. In general, while seeing an object as a visual scene it will be analysed by observing its features like edges, textures and colours. Similarly an audio/sound scene can be analysed by hearing perception. This concept is known as auditory scene analysis (ASA) (Wang and Brown, 2006). The field of ASA develops a computational method for analysing audio streams from various environments. Machine learning systems have more difficulties in finding the perception of human auditory systems in realistic acoustic scenes. The recent progress on machine learning and signal processing has many developments for automatic analysis of audio scenes and events.

Recently, everyday sound in real life environments which are called as domestic sounds pays more research attention. This emerging research area is known as computational auditory scene analysis (CASA). The motive of CASA is to understand unpredictable sound mixtures of day-to-day life including non-speech and music. CASA can be used for automatic speech recognition in real world environments. Listening to music focus on perceptual characteristics such as pitch and loudness of the sound, whereas listening to domestic sounds focus on the events present to learn the relevant information of the sound in order to categorise the environment. Research on categorising everyday sound is based on the attributes related to sources, actions and contexts. This kind of representation provides an acceptable generalisation and reduces misclassification.

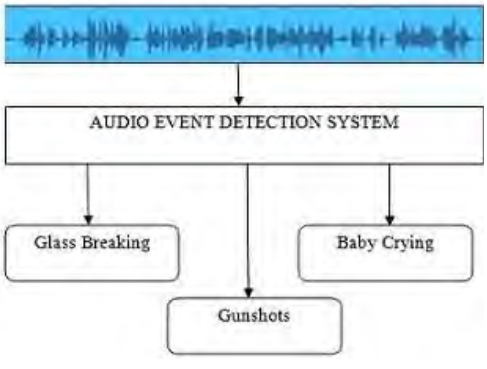
Computational analyses for speech and music signals are used in various real time applications. Speech applications include *speaker recognition* which recognises and identify the person who is talking, *automatic speech recognition* for recognising sequence of words in speech. Music applications are collectively called as *music information retrieval* (MIR) includes *automatic music transcription* to recognise the musical instruments played, genre classification identifies the style of music album. Nowadays non-speech sound scenes and events are getting more attention from the researchers. A real time auditory signal will be a mixture of background noise along with the source sound. This makes the auditory system complex to segregate the noise and perceive only the target sound event. So the computational analyses are completely based on machine

learning techniques, because the system will learn the parameters automatically with the previous knowledge of target sound events.

Sound event provides the rich information about the events taking place in the surroundings. System senses what action is happening in source sound and describes a meaningful context to listener in that environment. This is used in novel application areas such as public safety, security surveillances, activity detection, multimedia search, ambient assistive living, public healthcare monitoring, etc. Sound analysis system can be categorised into *classification* and *detection*. A process which automatically classifies a sound signal based on its characteristics into one of the categories/classes where the audio is recorded (example – office, home, metro station, and shopping mall) is called as audio scene classification. If the system addresses the information about localisation in time and labels with the events, then it performs *detection*.

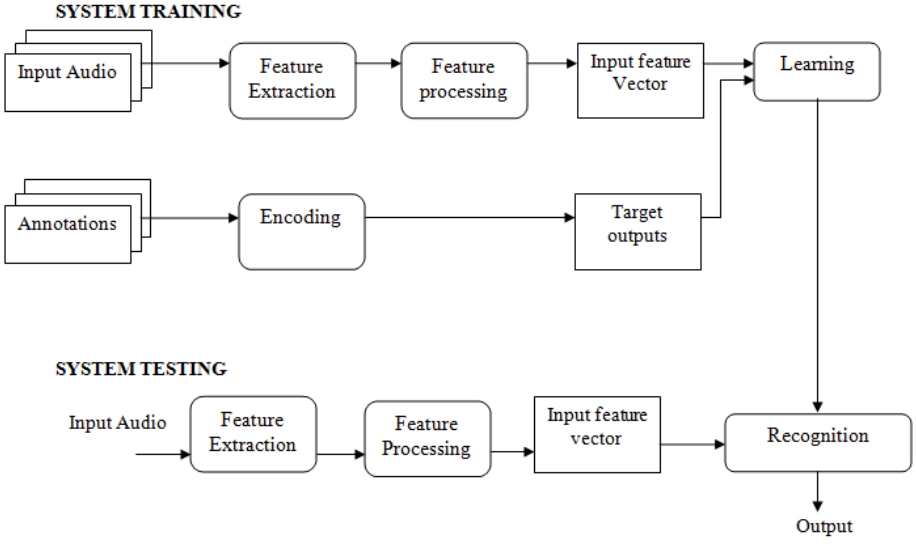
AED can be defined as analysing a continuous acoustic signal in order to extract the sound events present in the acoustic scene. Figure 1 illustrates overview of event detection. The system find the audio events in multisource conditions similar to our everyday life which includes speech, cough, laughter, baby cry, siren, street music, etc. Audio in real time environment includes overlapping of sound mixtures. AED aims to recognise the individual sounds in audio. It consists of two categories. The first one involves detection of audio events related in a particular scene which is called as polyphonic event detection (PED). The other one is recognising specific rare sound events in different environments called as monophonic event detection (MED).

Figure 1 Overview of audio event detection (see online version for colours)



The field of audio signal processing pays more attention to analyse these data (Alpaydin, 2014; Stowell et al., 2015). The computational system handling with real life environments are based on supervised machine learning techniques. Figure 2 illustrates the detailed architecture of an audio recognition system. Audio event detection is getting huge attention from research community in recent days due to its various advances in machine learning, internet of things (IoT) applications which include self-driving cars, acoustic sound monitoring in smart cities, etc. The proposed work uses deep learning model to build an audio event detection system to recognise various sound events in domestic environment.

Figure 2 Audio recognition system



The remaining of the paper is organised as follows. Related work in the field of audio event detection is reviewed in Section 2. Acoustic features relevant to audio event detection are summarised in Section 3. Various methodologies used for developing audio event detection are described in Section 4. Section 5 includes the detailed result and discussion. Finally, Section 6 concludes the paper.

2 Related works

Gencoglu et al. (2014) proposed a DNN classifier to recognise acoustic events. The DNN classifier obtained an accuracy of 60.3%, whereas the Hidden Markov Model (HMM) classifier yielded classification accuracy of 54.8%. McLoughlin et al. (2015) presented an acoustic events classification framework with spectrogram image-based front-end features, using support vector machine (SVM) and DNN classifiers. The results showed that the DNN-based classifier was superior to the SVM-based classifier under multiple signal-to-noise ratio (SNR) conditions. Niessen et al. (2013) proposed a two-layer hierarchical HMM for recognising acoustic events in which one layer for acoustic events and another layer for sub-event clusters. The results showed that the hierarchical HMM achieved an average frame-based F-measure score of 45.5%, and obtained better performance than the traditional GMM and HMM classifiers.

Clavel et al. (2005) proposed a novel detection approach to detect abnormal audio events for a surveillance system. The authors mainly considered the similarity between the acoustic features of variety of weapons by developing a hierarchical classification system. Giannoulis et al. (2016) proposed a dictionary selection method for an efficient background noise modelling which improves the conventional detection based on activation. Phan et al. (2016) proposed a discriminative decision forest to jointly learn the classification and regression. To perform classification from an overlapping sound mixture, the decision trees select the discriminative features to separate the positive event

segments from the negative ones. Regression is performed by letting the positive audio segments vote for onset and offsets of event occurrences.

Mesaros et al. (2011) a probabilistic latent semantic analysis (PLSA) system, a closely related approach to non-negative matrix factorisation (NMF), was proposed to detect overlapping sound events. Cotton and Ellis (2011) proposed a convolutive NMF algorithm applied on a Mel-frequency spectrum for detecting non-overlapping sound events. Elizalde et al. (2013) introduced I-vector system using GMM for speaker verification and scene detection. Agarwal et al. (2016) proposed a model using PLSA to monitor the co-occurrence of overlapping events. The model is used to learn the relationship between individual sound events in a polyphonic sound. Kroos and Plumbley (2017) addressed a neuroevolution of augmenting topologies (joint-NEAT) algorithm for sound event detection. The authors used wavelet-based deep scattering transform and k-means clustering for feature extraction.

Adavanne and Virtanen (2017) proposed a neural network architecture which learns relevant temporal information from the intermediate layers. This method generates strong labels for weakly labelled dataset. Cakir and Virtanen (2017) combined convolutional and recurrent layers as Convolutional Recurrent Neural Networks (CRNN) for rare sound event detection. Jeong et al. (2017) presented a CNN using short term and long term data in real life audio. The network learns the data better by applying optimisation strategies, adaptive thresholds, and class-wise early stopping. Lee et al. (2017b) described with sample level deep convolutional neural networks (DCNN) and multi scaled model on aggregating features for weakly supervised sound event detection. In proposed method, features are learned using multiple CNNs, multi level and multi scale feature aggregation, and final classification approaches. This outperforms comparing to spectrogram.

Kong et al. (2016) addressed supervised non-negative matrix factorisation (NMF) for separating noise from target events. Phan et al. (2017) proposed a system based on CNN and DNN coupled with novel weighted and multi-task loss functions and state-of-the-art phase-aware signal enhancement. The weighted loss is designed to tackle the common issue of imbalanced data in background/foreground classification while the multi-task loss enables the networks to simultaneously model the class distribution and the temporal structures of the target events for recognition. Grzeszick et al. (2017) presented a bag of features method for acoustic event detection and classification. The features are quantised for a supervised learning and histogram representation is computed. Schroder et al. (2017) evaluates a neural network system based on GMM, and HMM approaches for polyphonic acoustic event detection. The system is trained with amplitude modulation filter bank and Gabor filter bank.

Shuyang et al. (2017) proposed a method to optimise the sound event classification performance when labelling budget is limited and only a small portion of data can be annotated. The proposed method is called medoid-based active learning (MAL). K-medoids clustering is performed on sound segments, medoids are selected for labelling. Kumar and Raj (2017) propose a novel learning framework called Supervised and Weakly Supervised Learning (SWSL) where the goal is to learn simultaneously from weakly and strongly labelled data. The model is based on manifold regularisation on graphs in which we show that the unified learning can be formulated as a constraint optimisation problem which can be solved by iterative concave-convex procedure (CCCP). Zhang et al. (2015) extracted time frequency-based features by using tensor-based sparse approximation for audio event classification. Here discriminative features are extracted in spectro temporal domain used to classify events.

Lee et al. (2017a) proposed a sound event detection system that can recognise strong-labelled sound event from weakly-labelled data. It uses an ensemble of convolutional neural networks to detect audio events in the automotive environment. Each of the networks is based on various lengths of analysis windows for multiple input scaling. Cakir et al. (2015) found that using a set of independent detectors worked almost as well as a multilabel detector, and thus recommended the approach on the basis of its flexibility. Benetos et al. (2017) propose a method for tracking multiple overlapping sound events using linear dynamical systems which explicitly models the co-occurrence of sound event classes.

3 Acoustic features

3.1 Audio processing

Audio preprocessing is the first step to be carried out in machine learning before extracting acoustic features. Since the audio datasets are collected from various sources, they may have different sampling frequency. So it is necessary to resample into a uniform sampling frequency. In any sound recognition system the audio signals at higher frequency may also contain relevant information. In order to boost those higher energies these preprocessing in signal is preceded. The property of an audio/speech signal varies over time. Hence it is said to be non-stationary. To process the signal by digital means, it is necessary to sample the continuous-time signal into a discrete-time signal. The pre-processing steps are background noise removal, pre-emphasis, frame blocking, and windowing. So the signal is decomposed into number of frames to retain the signal for fixed duration and to extract the relevant information. Each frame is windowed to reduce the signal discontinuity. The next step to be carried out is feature extraction which transforms the audio signal in a compact representation.

3.2 Feature extraction

Any speech/audio signal can be represented as feature vector. The high level features are used in speaker recognition because of robustness and good performance over noisy environments. Low level spectral features are widely used as they are easy to compute. Features are more related to speech production mechanism and sound source filtering model. The process of extracting useful characteristics from any signal is called as feature extraction. It represents/transforms the signal in a more compact/numerical representation. The important aspect of feature extraction is reducing the redundant data in the signal. It makes the machine learning algorithm to function better because of reduced memory requirement and computational cost.

The audio features used in proposed work are summarised as follows:

- *Mel-frequency cepstral coefficients (MFCC):* MFCC are the short time power spectrum of any audio/speech signal which is derived as the linear cosine transform of log power spectrum on a nonlinear Mel scale frequency. For a windowed signal discrete Fourier transform (DFT) is computed using fast Fourier transform (FFT) algorithm to represent magnitude and phase in a signal. Human hearing perception of frequency bands is nonlinear. So Mel scale is approximated to linear frequency range

of below 1,000 Hz and logarithmic above 1,000 Hz. By applying these principles, Mel filter banks are computed. Each filter output is sum of its filtered spectral components. Logarithmic compression is mapped to magnitudes at each of Mel frequencies. Discrete cosine transform (DCT) is applied to Mel log amplitudes. Finally MFCCs are extracted as the resulting spectrum.

- *Log Mel spectrogram:* a spectrogram is a visual representation of audio signal. The spectrogram is a time frequency decomposition of a signal which implies its frequency content over time. Spectrogram can be computed by taking the magnitude of the short-time Fourier transform (STFT). Logarithmic compression to spectrograms is applied to square magnitude of output, because it compresses the dynamic range of values, i.e., low and high energy regions of the spectrum. To compute log-Mel spectrograms, Mel scale is considered as a reweighting of the frequency dimension which gives more relevant representation of the audio signal. In the proposed work log Mel spectrogram with 128 bands are extracted from the audio signal (44.1 kHz) with
  23 ms of window size and 50% of hop size.
- *Chromagram:* a chromagram of a short time Fourier transform is a twelve feature vector which is used to indicate energy level at each pitch classes in a signal. In audio retrieval-based applications, chroma features have proven to be a powerful mid level representation.
- *Spectral contrast:* spectral contrast feature could illustrate the relative distribution of the harmonic and non-harmonic components within the spectrum. FFT is applied on a signal to obtain the spectrum. Then, octave scale filters were applied to divide the frequency domain into sub bands. Here the strength of spectral peaks and their differences were estimated in each sub band. Logarithmic compression is applied. Finally,
  Karhunen-Loeve transform (K-L) is used to extract the spectral contrast feature. This feature provides more spectral information than MFCC.
- *Tonnez:* Tonnez-based representation estimates tonal centroids as coordinates in a six-dimensional interval space.

4 Methodology

AED aims to recognise the individual sounds in audio based on the temporal estimation of the events along with the labelled classes. Initially audio streams are decomposed into multiple frames. From each frame, a set of relevant audio features are extracted and represented as feature vectors. Feature vectors are given to proposed learning algorithm to learn the model. Gaussian mixture model (GMM)-based hidden Markov model (HMM) is widely used for AED. In proposed work, deep convolutional neural network (DCNN) is used to recognise AED. Deep architectures learn high level features from the low level features at each layer. Deep learning is one of the artificial intelligence (AI) which supports multilayer perceptron. Since deep learning can automatically learn the important features it is widely used in many domains like speech recognition, object detection, fraud detection, advanced search, classification, regression, etc.

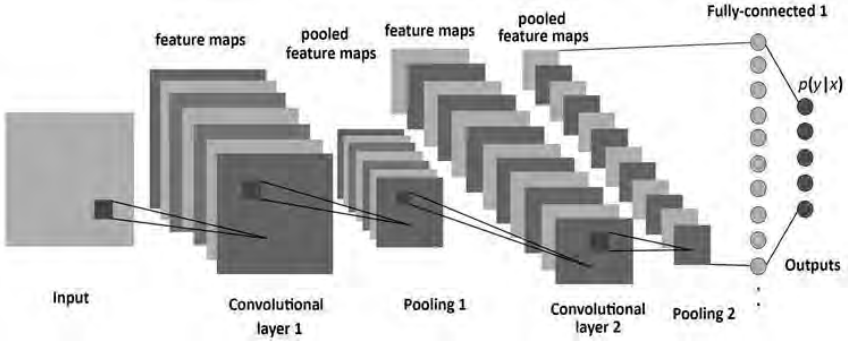
The objective of using deep architecture is that cognitive process is deep similar to humans. Deep learning supports both supervised and unsupervised learning techniques to learn the relationship associated with heterogeneous data. The availability of huge volume of raw data in real life makes a great demand to organise and analyse the data. Hence large scale data analysis helps to organise data and make the information more meaningful. Deep learning is one of the machine learning techniques. The traditional ANN with number of hidden layers is called as deep neural network (DNN). The advantage of deep learning is that it learns the features at various level of abstraction. Such that all intermediate features learned at each layer is also suitable for training.

Comparatively unsupervised data are widely used in deep learning for representing complex real time problems. The biggest challenge in big data is handling with real time streaming data. Training the deep learning model with big data is a tedious process since it requires large computational power. But this drawback has been significantly improved by the researchers with a large scale deep belief mechanism called as Large Scale Deep Belief Networks (DBN). The big data enriches the large scale DBN by supporting data parallelism and thus minimise the data transfer rate. The deep belief mechanism can process with millions of parameters, process them parallel in a convolution networks and store them in a global memory.

4.1 Convolutional neural network

Convolutional neural network (CNN) is similar to neural network and consisting of one or more convolutional layers and sub sampling layers followed by a fully connected neural network. The architecture of CNN is shown in Figure 3. The CNN organises the neurons and convolutions to make the network learn the features from the data.

Figure 3 Architecture of CNN



The convolution layer is an important part of CNN because it consists of set of filters. Each filter will be convolved independently with input data to map a feature vector. The filters can be randomly initialised through which a network can learn the parameters efficiently. Pooling is another concept of CNN used to reduce the spatial representation which helps to reduce the computation power of a network. Pooling layer is also called as down sampling layer.

5 Experimental setup

5.1 Datasets

The term sound scene refers to an environmental audio which represents the acoustic circumstances such as park, office, home, etc. Real life environment has different category of sound data and for each category, generation of sounds will be varied due to acoustic conditioning such as weather, number of people in a room, etc. This variability in training data will improve the generalisation in learning and the model can upgrade the previous learned algorithm by boosting which makes the machine learning on real time applications more effective and accurate. Datasets for the proposed work is collected from urban datasets. It consists of 8,732 sound clips each of 3 sec long with ten events such as air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music. The dataset by default is divided into 10-folds. Figure 4 shows the distribution of data.

Figure 4 Audio event labels (see online version for colours)

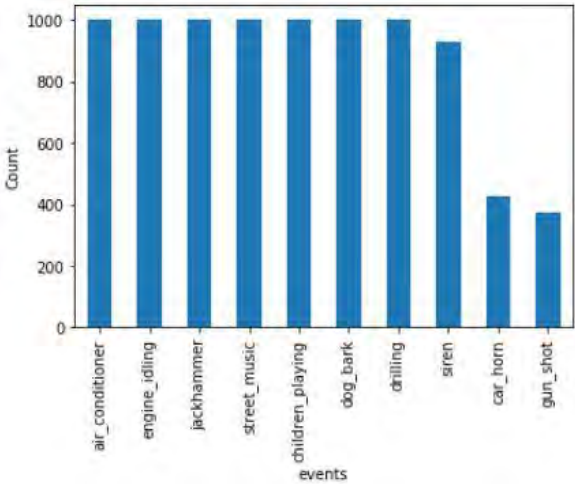
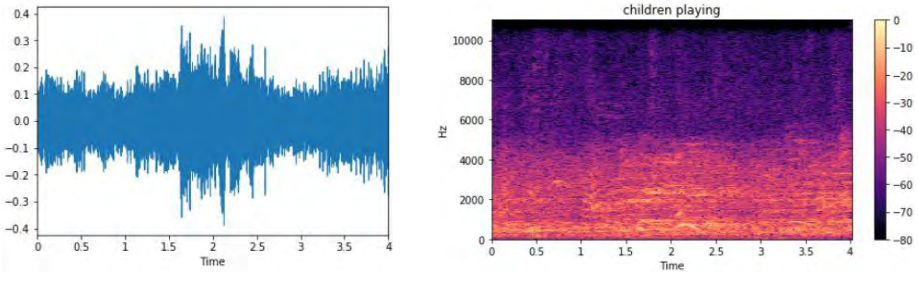
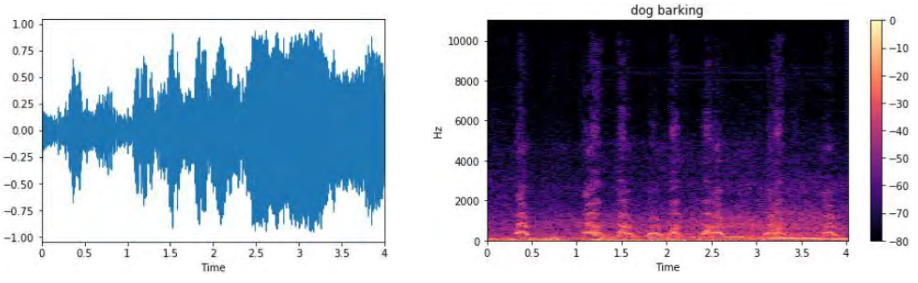


Figure 5 (a) Children playing (b) Spectrogram of children playing (see online version for colours)



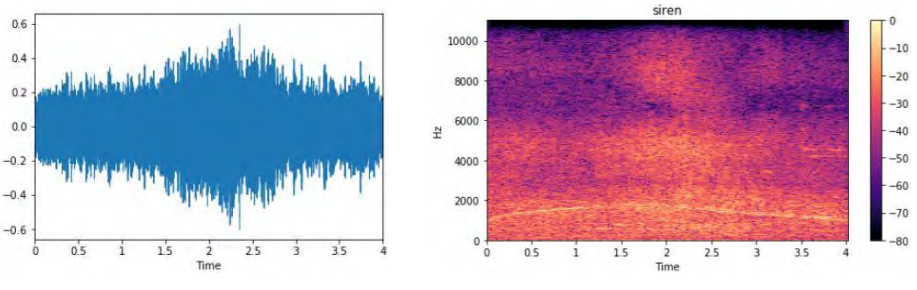
(a) (b)

Figure 6 (a) Dog barking (b) Spectrogram of dog barking (see online version for colours)



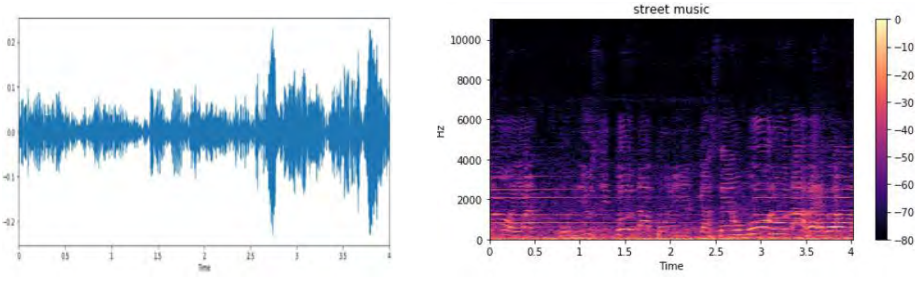
(a) (b)

Figure 7 (a) Siren (b) Spectrogram of siren (see online version for colours)



(a) (b)

Figure 8 (a) Street music (b) Spectrogram of street music (see online version for colours)



(a) (b)

Few of sound files are visualised with their spectrogram representations to understand how each sound is differing from other. Figure 5, 6, 7, and 8 shows the wave file plots of children playing, dog barking, siren and street music audio events and their spectrogram respectively.

The relevant features for the proposed work are extracted using Python Librosa library. For each input audio signal, features such as MFCC, log scaled Mel spectrogram, chroma, spectral contrast, and tonal centroids were extracted and stacked as feature vector. The proposed DCNN architecture consists of three convolutional layers. Various parameters of DCNN architecture is illustrated in Table 1, which shows how input is represented in the network. The input feature array is passed through convolution layers

and the output is pooled to reduce the dimension of data passed onto subsequent layers. Here, Conv2D layers are used for convolution operation. The number of filters in each convolutional layer will generate a feature map which extracts relevant features.

Filter size defines the height and width (rows and columns) of the filter kernel. The first layer performs convolution on the acoustic features over 24 filters with a kernel size of 5. Second and third layers also perform convolution similar to first layer over 48 filters with a kernel size 5. The filter size is increased to get higher level representation of data. A pooling operation determines the amount of data to be reduced from previous layer. The MaxPooling2D layers are used for max-pooling operation in the framework. This can speed up the training process and produce better results with signals. The network learns the input data with given parameters using nonlinear activation 'ReLU'. The feature vectors are extracted by convolutional filters and they are flattened into an array and passed into fully connected layer to recognise the audio events.

In order to avoid the model getting over fit, a powerful regularisation technique called dropout is used. Dropout enables the model to learn several independent representations of the same data by randomly deactivating neurons during the learning process. Other hyper parameter such as Batch Normalisation is used to normalise the input layer. This allows each layer in the network to learn independently from other layers. The model is trained initially for ten iterations with a batch size of 100. The number of iterations is increased based on the performance improvement. Finally, softmax activation is used to predict the audio events based on the probability. The proposed model is compiled using cross entropy as the loss function with Adam optimiser. The Adam optimiser is responsible for updating the neuron weights through back propagation.

Table 1 Parameters of proposed DCNN architecture

Layer	*Number of filters*	*Activation*	*Padding*
Input data	-		
Conv2D (f = 5, s = 1)	24	ReLU	-
Maxpool2D (f = 4, s = 2)	-		
Conv2D (f = 5, s = 1)	48	ReLU	Valid
Maxpool2D (f = 4, s = 2)	-		
Conv2D (f = 5)	48	ReLU	Valid
Flatten	-		
Dense (64)	-		
Softmax (10)	-		

5.2 Results and discussion

The proposed approach is evaluated in terms of audio event accuracy and F-score. The dataset sorted into ten stratified folds, and all models were evaluated using 10-fold cross validation. For training the proposed DCNN architecture, one of the nine training folds in each split is used as a validation set for identifying the training epoch that yields the best model parameters when training with the remaining eight folds. The performance of the proposed DCNN is illustrated in Tables 2a and 2b.

Table 2a Performance for each event

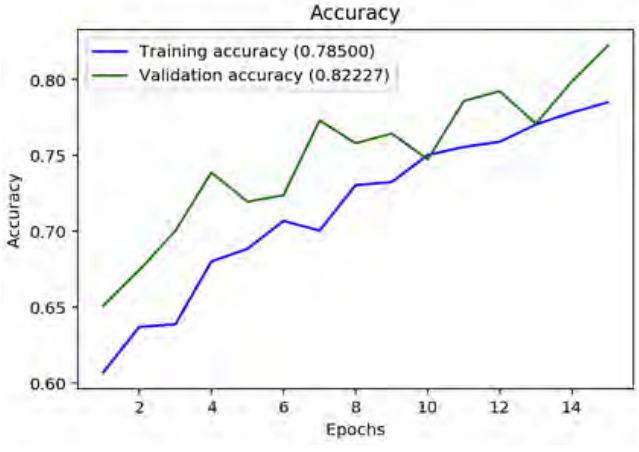
Event name	*F-score*
Air conditioner	0.78
Car horn	0.91
Children playing	0.77
Dog bark	0.80
Drilling	0.88
Engine idling	0.85
Gunshot	0.82
Jackhammer	0.80
Siren	0.82
Street music	0.82

Table 2b Performance metrics report

Model	*F-score*
DCNN + log Mel spectrogram	0.82

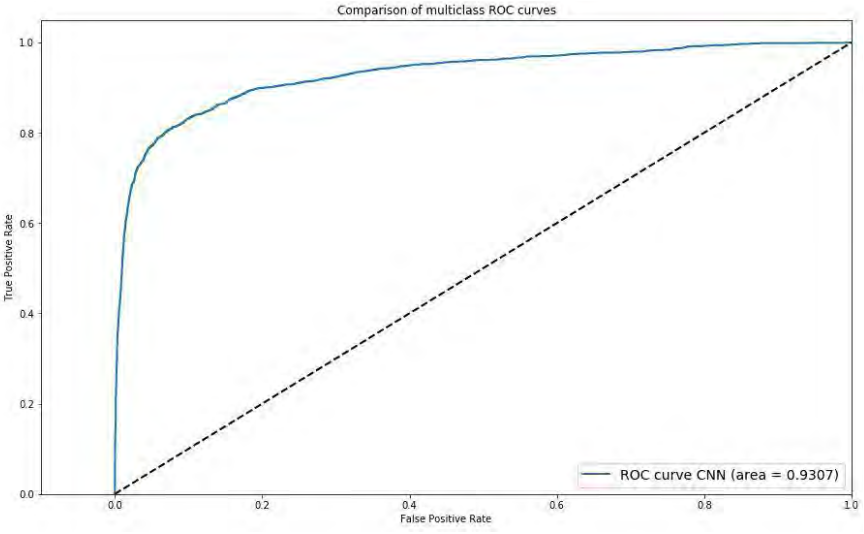
The accuracy plot shown in Figure 9 illustrates the accuracy on train and validation datasets. The result shows a progress in accuracy which indicates that, the model has the ability to generalise the novel data.

Figure 9 Accuracy for DCNN (see online version for colours)



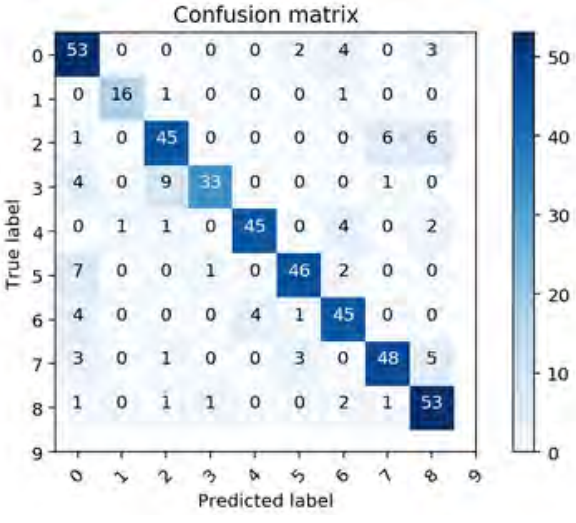
Receiver Operating Characteristics (ROC) for the proposed DCNN is illustrated in Figure 10 represents true positive rate plotted with false positive rate with different cutoff points. ROC plot evaluates the classifier which predicts the rare sound events.

Figure 10 ROC curve for DCNN (see online version for colours)



Confusion matrix shown in Figure 11 determines the model prediction. The horizontal axis shows the predicted classes and the vertical axis the actual classes.

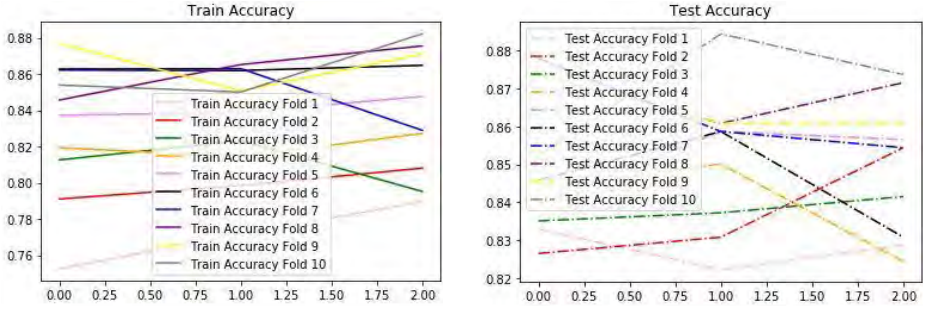
Figure 11 Confusion matrix for proposed DCNN (see online version for colours)



Events: air conditioner (0), car horn (1), children playing (2), dog bark (3), drilling(4), engine idling(5), gunshot (6), jackhammer (7), siren (8), street music (9).

The performance of 10 folds cross validation is shown in Figure 12 and it proves that during test process fold 8 and fold 10 performs better than other cross folds. The overall score of proposed model with 10-fold cross validation is 84%

Figure 12 (a) Train accuracy for 10 fold (b) Test accuracy for 10 fold (see online version for colours)



(a) (b)

The k fold scores of each fold are shown in Table 3.

Table 3 K fold score

Folds	*Score*
Fold 1	0.82
Fold 2	0.85
Fold 3	0.84
Fold 4	0.82
Fold 5	0.85
Fold 6	0.83
Fold 7	0.85
Fold 8	0.87
Fold 9	0.86
Fold 10	0.87

6 Conclusions

In this work, audio event detection is proposed based on a DCNN framework. The proposed DCNN framework handles multiple features stacked and process feature maps which make the network to learn sound events better. The performance of the system was evaluated with urban dataset which proves that log Mel band spectrogram features improved the event detection with an F-score of 0.82. The proposed model evaluated with 10-fold cross validation further improved the performance with 84% of accuracy. The future progress in the field will be concentrated with new larger datasets, processed with Apache Spark framework to improve the effectiveness of large scale data using deep learning techniques.

References

Adavanne, S. and Virtanen, T. (2017) 'Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network', *IEEE Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Agarwal, A., Quadri, S.M., Murthy, S. and Sitaram, D. (2016) 'Minimally supervised sound event detection using a neural network', *ICACCI*.

Alpaydin, E. (2014) *Introduction to Machine Learning*, 3rd ed., MIT Press.

Benetos, E., Lafay, G., Lagrange, M. and Plumbley, M.D. (2017) 'Polyphonic sound event tracking using linear dynamical systems', *IEEE/ACM Trans. Audio Speech Lang. Process*, Vol. 25, No. 6, pp.1266–1277.

Cakir, E. and Virtanen, T. (2017) 'Convolutional recurrent neural networks for rare sound event detection', *IEEE Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Cakir, E., Heittola, T., Huttunen, H. and Virtanen, T. (2015) 'Multi-label vs. combined single-label sound event detection with deep neural networks', *23rd European Signal Processing Conference (EUSIPCO)*, pp.2551–2555.

Clavel, C., Ehrette, T. and Richard, G. (2005) *Events Detection for an Audio-Based Surveillance System*, p.1–4, Published by IEEE.

Cotton, C.V. and Ellis, D.P.W. (2011) 'Spectral vs. spectro-temporal features for acoustic event detection', *Proc. IEEE Workshop Appl. of Signal Process. Audio Acoust.*, pp.69–72.

Elizalde, B., Lei, H., Friedland, G. and Peters, N. (2013) 'An I-vector based approach for audio scene detection', *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Gencoglu, O., Virtanen, T. and Huttunen, H. (2014) 'Recognition of acoustic events using deep neural networks', *The 22nd European Conference on Signal Processing*, pp.506–510.

Giannoulis, P., Potamianos, G., Maragos, P. and Katsamanis, A. (2016) 'Improved dictionary selection and detection schemes in sparse-CNMF-based overlapping acoustic event detection', *IEEE Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Grzeszick, R., Plinge, H. and Fink, G.A. (2017) 'Bag-of-Features methods for acoustic event detection and classification', *IEEE/ACM Transactions on Audio, Speech, And Language Processing*, Vol. 25, No. 6, pp.1242–1252.

Jeong, I.Y., Lee, S., Han, Y. and Le, K. (2017) 'Audio event detection using multiple-input convolutional neural network', *IEEE Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Kong, Q., Sobieraj, I., Wang, W. and Plumbley, M.D. (2016) 'Deep neural network baseline for DCASE Challenge 2016', *IEEE Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Kroos, C. and Plumbley, M.D. (2017) 'Neuroevolution for sound event detection in real life audio', *IEEE Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Kumar, A. and Raj, B. (2017) 'Audio event and scene recognition: a unified approach using strongly and weakly labeled data', *IEEE Transaction*, IEEE Transaction, pp.3475–3482.

Lee, D., Lee, S., Han, Y. and Lee, K. (2017a) 'Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input', *IEEE Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Lee, J., Park, J., Kum, S., Jeong, Y. and Nam, J. (2017b) 'Combininng multi-scale features using sample-level deep convolutional neural networks for weakly supervised sound event detection', *IEEE Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*.

McLoughlin, I., Zhang, H. and Xie, Z. (2015) 'Robust sound event classification using deep neural networks', *IEEE/ACM Trans. Audio Speech Lang. Process.*, Vol. 23, No. 3, pp.540–552.

Mesaros, A., Heittola, T. and Klapuri, A. (2011) 'Latent semantic analysis in sound event detection', *Proc. Eur. Signal Process. Conf.*, pp.1307–1311.

Niessen, M.E., Kasteren, T.L.M.V. and Merentitis, A. (2013) 'Hierarchical modeling using automated sub-clustering for sound event recognition', *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.1–4.

Phan, H., Becker, M.K., Gerkmann, T. and Mertins, A. (2017) 'DNN and CNN with weighted and multi-task loss functions for audio event detection', *IEEE Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE2017)*.

Phan, H., Hertel, L., Maass, M., Koch, P. and Mertins, A. (2016) 'CaR-Forest: joint classification-regression decision forests for overlapping audio event detection', *IEEE Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Schroder, J., Moritz, N., Anemuller, J., Goetze, S. and Kollmeier, B. (2017) 'Classifier architectures for acoustic scenes and events: implications for DNNs, TDNNs, and perceptual features from DCASE 2016', *IEEE/ACM Transactions on Audio, Speech, And Language Processing*, Vol. 25, No. 6, pp.1304–1314.

Shuyang, Z., Heittola, T. and Virtanen, T. (2017) 'Active learning for sound event classification by clustering unlabeled data', *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.1–5.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M. and Plumbley, M.D. (2015) 'Detection and classification of acoustic scenes and events', *IEEE Transactions on Multimedia*, Vol. 17, No. 10, pp.1733–1746.

Wang, D.L. and Brown, G.J. (2006) *Fundamentals of Computational Auditory scene Analysis: Principles, Algorithms, and Applications*, Book Review, Published by IEEE.

Zhang, X., He, Q. and Feng, X. (2015) 'Acoustic feature extraction by tensor based sparse representation for sound effects classification', *IEEE ICASSP*, pp.166–170.