Automatic Audio Event Recognition Schemes for Context-Aware Audio Computing Devices

Thesis submitted to the

Indian Institute of Technology Bhubaneswar

For award of the degree

of

Master of Technology

by

Shivam Soni

(17EC06014)

Under the guidance of

Dr. M. Sabarimalai Manikandan

SCHOOL OF ELECTRICAL SCIENCES

INDIAN INSTITUTE OF TECHNOLOGY BHUBANESWAR

APRIL 2019

APPROVAL OF THE VIVA-VOCE BOARD

26/04/2019

Certified that the report entitled **"Automatic Audio Event Recognition Schemes for Context-Aware Audio Computing Devices"** submitted by **Shivam Soni** (17EC06014) to the Indian Institute of Technology Bhubaneswar in partial fulfillment of the requirements for the degree Master of Technology has been accepted by the examiners during the viva-voce examination held today.

(Supervisor) **(External Examiner)**

(Internal Examiner 1) **(Internal Examiner 2)**

CERTIFICATE

This is to certify that the report entitled **"Automatic Audio Event Recognition Schemes for Context-Aware Audio Computing Devices"** submitted by **Shivam Soni** (17EC06014) to Indian Institute of Technology Bhubaneswar is a record of bonafide research work under my supervision and the report is submitted for final evaluation of the M. Tech thesis report.

Date :

Dr. M. Sabarimalai Manikandan
Assistant Professor
School of Electrical Sciences
Indian Institute of Technology Bhubaneswar
Bhubaneswar, India

DECLARATION

I certify that

a. The work contained in the thesis is original and has been done by myself under the general supervision of my supervisor.

b. The work has not been submitted to any other institute for any degree or diploma.

c. I have followed the guidelines provided by the institute in writing the thesis.

d. I have conformed to the norms and guidelines given in the ethical code of conduct of the institute.

e. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

f. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Shivam Soni

Acknowledgments

First, I would like to thank my supervisor, Dr. M. Sabarimalai Manikandan, for being great teacher and providing guidance for me over the past one year, and for his great support and encouragement to pursue my own ideas. I am really grateful to has been his student. I want to thank all the current and past IIT Bhubaneswar professors and students for the great and fruitful environment they created.

I am grateful for my seniors Vadrevu Simhadri and G. Narendra Kumar Reddy for their help and valuable comments and discussions. I am indebted to Gaurav and Lipsa Routray for their support for my work in its earliest stage. I would like to thanks each member of biomedical signal processing lab for all their help and insights throughout my M.Tech project. Especially, I want to thanks Swaraj Mohapatra, Kinjarapu Manoj Kumar, Sudhir Kumar Sahoo, Sudipta Dey, Saurabh Arkasali, M. Sai Raghvendra Dinesh, Vatsalya Chaubey and Karinki Manikanta for their support.

I would like to thank my M.Tech seniors, Ishank Goel, Taufeeq Ahmed, Gaurav, and Parul for their great support. I owe many thanks to my colleagues Zore Tukaram Chandrakant, Manmeet Singh, Mirza Dawood Ali, Deepak Kumar, G. Shruti, Aditya Sethi, Shruti, Pooja Kumawat, Krishna Nandanam, Vishal Sherkhane, Mukesh Kumar, Lankipalli Vinod Kumar and Moganti Kiran Kumar for motivating me all the time.

But most of all, I want to express my deepest gratitude to my family who supported me in every possible way. Especially, I want to thanks Shanky, Aman, Sufiyan, Salim, Harsh, Ram, Shashikant, Shivam and Anjali Rastogi for their support and love all the time.

Shivam Soni

Abstract

Automatic audio event recognition (AER) plays a major role in designing and building intelligent-location and context-aware applications including audio surveillance, audio indexing and content retrieval, drone and robotic navigation, machine health monitoring, audio-aware voice processing services, and urban sound pollution monitoring. In this thesis, we present audio event recognition schemes using two popular audio features Mel-frequency cepstral coefficients (MFCC) and spectrogram image feature (SIF) and machine learning and deep learning based classifiers such as multi-class support vectors machines (MC-SVM), feed-forward neural networks (FFNNs), one-dimensional convolutional neural networks (1D-CNNs) and two-dimensional convolutional neural networks (2D-CNNs) that are capable of automatically recognizing seven sound classes including *aircraft, construction, music, nature (wind and rain), speech, vehicle, and train.* In this study, we created large scale audio database for both training and testing purposes. The performance of the four AER schemes are evaluated under different audio frame sizes (100 ms, 250 ms and 500 ms) using a wide variety of sounds recorded using different kinds of recording devices. In the first stage of experiment, we used MFCC feature to represent the audio and three classifiers MC-SVM, FFNN and 1D-CNN to detect audio events. The results shows that the FFNN and 1D-CNN based AER schemes had the F1-score values of 95.72% and 96.34% for audio frame size of 250 ms whereas MC-SVM based AER scheme had the F1-score values of 85.84%. The 1D-CNN based AER scheme had a class-wise accuracy is greater than 84% for audio frame size of 250 ms whereas the FFNN based scheme had a class-wise accuracy is greater than 80% for audio frame size of 250 ms. In the second stage of experiment, we used spectrogram image feature to represent audio signal and image classifier 2D-CNN to detect the events. The result showed that 2D-CNN based AER scheme had the F1-score value of 87.04% and overall accuracy of 88.48% for audio frame size of 100 ms. In this 2D-CNN based AER scheme, the class-wise accuracy for 100 ms frame size is greater than 250 ms and 500 ms frame size. All four AER schemes were evaluated in real time by using our audio event monitoring graphical user interface.

Contents

List of Figures

List of Tables

List of Abbreviations

KNN	K- Nearest Neighbors
LVQ	Learning vector quantization
ANN	Artificial neural networks
GMM	Gaussian mixture models
HMM	Hidden Markov models
STFT	Short-time Fourier transform
FWT	Fast wavelet transform
CWT	Continuous wavelet transform
DTW	Dynamic time warping
LTS	Long-term statistics
HCC	Homomorphic cepstral coefficients
TDMFCC	Two dimensional mel frequency cepstral coefficients
DTDMFCC	Dynamic two dimensional mel frequency cepstral coefficients
LDA	Linear Discriminant Analysis
PCA	Principal Component Analysis
MP-TFD	Matching-pursuit time-frequency distribution
NMF	Non-negative matrix factorization
SAI	Stabilized auditory image
DCT	Discrete Cosine Transform
PLPC	Perceptual linear prediction coefficients
LPCC	Linear predictive cepstral coefficients
DWC	Discrete wavelet coefficients
MFDWC	Mel filter-bank discrete wavelet coefficients
ZCR	Zero Crossing Rate
LBP	Local Binary Pattern

List of Symbols

exp$(\cdot)$	Exponential
SR	Sampling rate
b/s	bits per sample
h	Hours
min	Minute
ER	Error
BD	Bit depth
A	Aircraft
C	Construction
M	Music
N	Nature
S	Speech
T	Train
V	Vehicle

Chapter 1

Introduction

In last few decades, humankind saw a quick difference in day by day way of life with the introduction of computers in our lives. When we initially experienced with the computers, the vast majority of us felt that they are excessively unpredictable, excessively massive and they won't be extraordinary use for individuals aside from researchers. With the assistance of the fast advancements in hardware, computers got smaller and simpler to utilize. Thus, they began to turn into a central point in our everyday lives.

The more people become used to the computers, the more they began asking from them. The unnatural bond between people and an electronic gadget shrinked quickly as of late. The general population of the 21st century are comfortable with conveying electronic devices with them anyplace they go to. These days, we anticipate that computers should enable us to understand our world surroundings and know about what is happening. We need electronic gadgets that can percept the physical context and help their clients in specific jobs, without any direction. Consequently, setting context-aware gadgets have been a mainstream look into region among the AI researchers. A portion of the applications, for example, navigation assistance as of now partake in our every day lives, yet they give just constrained context-awareness [5].

In fact, designing context-aware devices is simpler said than done. People combine visual and sound information for view of environment, yet this isn't in any case the

case for AI frameworks.

Table 1.1: Environmental sound and its application.

Audio Events	**Audio Scene Environments**	**Outdoor/Indoor Scenarios**	**Some of Applications**
Fan Sound	Living	Indoor	Machine Fault Monitoring, Context-Aware Audio Processing
Air-conditioner Sound	Living	Indoor	
Train Sound	Vehicle, Travelling, Living	Outdoor/Indoor	Engine Fault Monitoring, Context-Aware Audio Processing, Urban Traffic Tracking, Sound Pollution
Four-Wheeler Sound	Vehicle, Travelling, Living	Outdoor/Indoor	
Two-Wheeler Sound	Vehicle, Travelling, Living	Outdoor/Indoor	
Auto-Rickshaw Sound	Vehicle, Travelling, Living	Outdoor/Indoor	
Vehicle Horn Sound	Vehicle, Travelling, Living	Outdoor/Indoor	Urban Traffic Tracking, Sound Pollution Monitoring
Aircraft/Helicopter Sound	Vehicle, Travelling, Living	Outdoor	Engine Fault Monitoring, Context-Aware Audio Processing
Drone/Quadcopter Sound	Vehicle	Outdoor	
Boat Sound	Vehicle, Travelling	Outdoor	Engine Fault Monitoring, Costal Surveillance
Ship Sound	Vehicle, Travelling	Outdoor	
Wind Sound (whistling, gusting, rustling)	Forest, Travelling, Building	Outdoor/Indoor	Context-Aware Audio Processing
Rain Sound	Forest, Travelling, Building	Outdoor/Indoor	
Bird Sound	Forest/Farm, Travelling, Building	Outdoor/Indoor	Context-Aware Surveillance
Applause Sound	Meeting and Entertainment	Outdoor/Indoor	Multimedia Indexing
Crowd Cheers	Meeting and Entertainment	Outdoor/Indoor	Multimedia Indexing
Laughter Sound	Meeting and Entertainment	Outdoor/Indoor	Multimedia Indexing
Speech Sound	Meeting and Entertainment	Outdoor/Indoor	Multimedia Indexing
Music Sound	Entertainment	Outdoor/Indoor	Multimedia Indexing

These days, by far most of these gadgets are exclusively based on processing visual data. Despite the fact that it is conceivable to get exact, precise details of the surroundings from the picture information, physical difficulties exist. For the applications, for example, camera based indoor robot navigation even a little loss of sight would cause

difficult issues.

On the other hand, sound data can give extra attention to these frameworks. One of the senses that we utilize most while interacting with the world is hearing. However, computers still fail badly, compared with people, with regards to translating and assigning meanings to the sounds. With the expansion of accessible sound information and handling power, another way opened for research and improvement for the researchers [35].

1.1 Background and Motivation

Sounds convey lots of information about our everyday environments that can be used for extracting interesting portion of the events present in the recorded audio and video and predicting the faults different kinds of machines that are continuously used in many practical applications [22]- [35]. Table 1.1 shows the some environmental sounds and its applications in real world scenario. Further, the sounds can be used for monitoring some of suspicious activities in the unauthorized zones. The issue of recognizing risky occasions on streets or roads was considered by designing an audio surveillance framework that automatically identifies dangerous circumstances, for example, vehicle crashes and tire slipping [12]. Audio event recognition (AER) is defined as the detection of individual sound events that are available inside an audio, resulting in a symbolic description such that each annotation gives the start time, end time and label for a single instance of a specific sound event [35]. Audio scene classification (ASC) is to classify a test recording into one of predefined sound scenes that characterizes the everyday environment in which a recording has been made, on the assumption that an audio scene is distinguishable from others based on its general acoustic properties [52]. Recently, ASC is a prominent research topic which is considered as a machine-learning task within a widespread single class classification paradigm, wherein a set of class labels is known. The ASC system can select exactly one for any given input. Automatic audio event detection and scene recognition in real-life audio presents many difficulties

such as the inherent acoustic variability of the sounds belonging to the same sound event class, or other sounds overlapping with the sound event of interest [46]. Further, the audio signal is more versatile when audio events are mixed with one or more background sound sources with different mixing ratios. However, a significant amount of research is still needed to reliably recognize individual sound in real-life soundscapes, where multiple sound sources are present, often simultaneously, and distorted by the environment and recording instruments [51].

Our inspiration to handle this issue originates from the yet unsatisfactory outcomes in this area. With the selection of right features and by utilizing the ongoing upgrades in the AI, for example, deep learning, we trust that high accuracy event recognition is conceivable, in real-life scenario.

1.2 Major Goals

In this thesis, we present audio event recognition (AER) schemes using the mel-frequency cepstral coefficients (MFCC) and spectrogram image as features and machine classifiers such as multi-class support vectors machines (MC-SVM), feed-forward neural networks (FFNNs), one-dimensional convolutional neural networks (1D-CNNs) and two-dimensional convolutional neural networks (2D-CNNs) that are capable of automatically recognizing seven sound classes including aircraft, construction, music, nature (wind and rain), speech, vehicle, and train sounds. Major contributions of this thesis work are as follows.

- Exploring an effective deep learning networks based automatic audio event recognition scheme for audio surveillance application.
- Identifying the optimal audio frame size for yielding better recognition accuracy with considerable computational load under real-time implementation of audio surveillance system.

- Creating the large scale audio databases including aircraft, construction, music, nature (wind and rain), speech, vehicle, and train sounds by considering different audio recording devices such as commercial hand-held audio recorders, Smartphones, laptop PC and different recording environments.
- Developing cloud-based audio event recognition (AER) framework for remote audio sensing and monitoring services.

1.3 Organization

The rest of this thesis is organized as follows. Chapter 2 presents the literature review related to this work .Chapter 3 presents audio event recognition (AER) schemes using the MFCC and spectrogram as features using machine learning and deep learning based classifiers. Chapter 4 presents the evaluation results of the four AER schemes on large scale audio databases. Finally, conclusions and future directions are given in Chapter 5.

Chapter 2

Literature Survey

In this chapter, we described the summary of each technique, various features for audio event recognition, various classifier, real time application and motivation behind this thesis work.

2.1 Conventional Machine Learning Based Methods

In this section, we described the brief overview of AER schemes based on conventional machine learning classifiers such as KNN, GMM, SVM, HMM, LDA, MLE, ANN from literarure. In [2], Wang et al., proposed a method to classify the environmental sound using hybrid classifier (combination of KNN and SVM). The technique was evaluated for 12 class sound using spectral domain feature. The results showed that hybrid classifier perform better than HMM for the same dataset. In [1], Goldhor presented environment sound recognition (ESR) based on two dimensional cepstral coefficient feature and maximum likelihood estimation (MLE) classifier. The method was evaluated for 23 class sounds which is recorded by putting the recorder at different orientation and distance. The results showed that the accuracy of near field recorded sound is more than

far field recorded sound. In [3], Cowling et al., presented a comparative study using various classifiers (ANN, LVQ, GMM, LTS, DTW) for environmental sound classification. The schemes were evaluated on 8 class environmental sound with spectral and cepstral domain feature. The results showed that DTW with MFCC feature is superior than other classification technique. In [6], Souli et al., proposed audio sounds classification using scattering features and support vectors machines for medical surveillance applications. In [7], Hwang et al., proposed KNN based method to classify environment audio recognition for smart mobile application. The method was evaluated for three types of audio datasets i.e., audio scene, audio events and phone context based on Gaussian histogram features. The results showed that the recognition rate of phone context is more than others. In [8], Valero et al. presented the biologically inspired Gammatone cepstral coefficients (GTCCs) feature based non speech recognition scheme with SVM classifier. The results showed that the GTCCs feature results in better classification accuracy than that of the MFCC based scheme. In [9], Lee et al., presented a method for classification of birds into different species based on KNN classifier. The method was evaluated for a dataset of 28 bird species with TDMFCC and DTDMFCC as feature. The results showed the classification accuracy of 84.06%. In [10], Ghoraani et al., proposed method for environment sound classification based on LDA classifier. The method was evaluated for 10 class sound with time-frequency matrix (TFM) features. The results showed that the TFM feature results in better classification accuracy than that of the MFCC based scheme. In [11], Lin et al., presented a technique to categorize and classify the audio events using multiclass support vector machines (MC-SVMs) as a classifier. The method is evaluated for a public audio database Muscle Fish consist of 16 class sounds with spectral domain features. The results showed the 91-97% classification accuracy. In [12] Almaadeed et al., proposed SVM classifier based an audio surveillance system to support visual surveillance system for detecting hazardous event on roads such as car crashes and tire skidding. The method was evaluated for MIVIA road dataset with temporal and spectral features. In [13], Chu et al., presented GMM and KNN classifier based environment sound classification using matching pursuit(MP)features. The method was evaluated on 14 class environmental sound. The

results showed that MP features results in better classification accuracy than that of the MFCC based scheme. In [14], Chu et al., proposed a method to recognize environmental sound using KNN and GMM as classifier. The method was evaluated on 14 class environmental sound with MP algorithm base time-frequency features. The results showed that the classification accuracy of KNN is more than GMM classifier. In [16], Wang et al., presented environment sound classification for home automation using Gabor dictionary with matching pursuit (MP), principal component analysis (PCA) and linear discriminate analysis and multi-class support vector machine (MC-SVM) classifier for 17 class sounds recognition. In [17], Wang et al., proposed sound recognition for smart home using MC-SVM classifier. The method was evaluated for 10 class environment sound with wavelet subspace based features. The results showed that classification accuracy of this method is good under noise condition. In [19], Tran et al., proposed an online sound recognition based on probabilistic distance SVM with subband temporal envelop feature. The method was evaluated using the 10 class sound database. Results showed that the method outperforms than the MFCC features with SVM classifier. In [21], Yang et al. showed that spectral MFCC features along with temporal dynamic feature can improve the performance for sound classification with local binary pattern. The method is evaluated on the TUT database (15 class) with proposed features with an ensemble classifier. In [22] Rabaoui et al., proposed one class SVM based environmental sound classification for audio surveillance of 9 sound classes using the audio features such as zero-crossing rate (ZCR), spectral centroid, spectral bandwidth, MFCCs extracted in various domains including temporal, frequency, cepstral, and wavelet. The results showed that it outperforms the binary class SVM and HMM based schemes. In [25], Souli et al., proposed a method for environmental sound classification using MC-SVMs as a classifier. The method was evaluated on 10 class audio dataset using magnitude spectrogram feature. The reults showed the recognition accuracy of 89.62%.

2.2 Deep Learning Based Methods

In this section, we described the brief overview of AER schemes based on deep learning classifiers such as FFNN, 1D-CNN, RNN, 2D-CNN from literarure. In [4] Beritelli et al., presented environmental sound classification using the MFCC features and feed forward neural network classifier. The scheme was evaluated on 10 sound classes by varying window size. In [5], Cakir et al. presented a convolutional recurrent neural network (CRNN) based polyphonic sound recognition scheme with MFCC feature. The results showed that the RNN based scheme considerably improves the recognition accuracy as compared to the feed forward neural networks (FFNNs), convolutional neural networks (CNNs) and recurrent neural networks (RNNs) based schemes on TUT database. In [18], Rakotomamonjy, proposed the use of supervised feature learning approaches for extracting relevant and discriminative features from acoustic scene recordings. The method was evaluated on DCASE-16 and Rouen -15 dataset. The result showed that for smaller scale dataset, SNMF classifier is slightly less prone to over-fitting than convolutional neural networks. In [20], McLoughlin et al.,proposed machine learning technique (SVM, DNN) to classify the environmental sound. This methods was evaluated for two types of public audio datasets (RWCP, NOISEX-92) with time-frequency features (SAI, SIF). The result showed that the classification accuracy of DNN classifier with SIF feature was more than others. In [23], Sigtia et al. presented a comparative study using various classifiers (DNN, GMM, SVM, RNN) in the context of IoT sound sensing application. The schemes were evaluated on the DCASE-16 database with 13 class of sounds. The results showed that the DNN yields the best classification accuracy than the GMM and SVM based schemes for a range of computational costs. In [24], Gencoglue et al. proposed the deep neural network (DNN) based recognition of isolated acoustic events such as footsteps, baby crying, motorcycle, rain etc. The scheme was evaluated for 61 sound classes by using the Mel energy feature vector. The results demonstrated that the DNN based audio classifier performs better than the HMM with GMM classifiers. In [26], Zhang et al., presented deep learning (2D-CNN) based audio event recognition. The method was evaluated on

two types of standard datasets (RWCP , NOISEX-92) with spectrogram as feature. The results showed the 93% classification accuracy. In [27], Boddapati et al., proposed that CNN classifier (AlexNet, GoogleNet) which is designed specially for object identification in images, can be successfully classify spectral image of environmental sound. The method was evaluated on standard audio datasets (ESC-10, ESC-50, Urbansound8-K) using MFCC, Spectrogram and Chorma based feature(CRP). The results showed that recognition rate in Urbansound8-K dataset is more than others. In [30] Phan et al., proposed an approach for acoustic scene classification using deep GRU based Recurrent neural network (RNNs). The experiment was conducted on largest available dataset i.e., LITIS Rouean (19 scene, 1500 minute). The results showed that the error rate is less in RNN classifier compare to other DNN classifier.

Table 2.1 and Table 2.2 showed the summary of AED schemes and summary of standard audio database respectively.

Table 2.1: Literature review summary of audio event recognition.

Author	**Database**	**Features**	**Classifier**	**Accuracy**	**Remarks**
Wang et al., [2]	12class, f_s = $16kHz$, r=16 bit/sample	Spectral centroid, spectral flatness, spectral spread	Hybrid Classifier (KNN & SVM)	85.1%	Frame level approach for ESC using hybrid classifier
Goldhor, [1]	23class, f_s = $16kHz$, r=14 bit/sample	Cepstral Coefficients	Maximum Likelihood Estimation	Far field (93%), Near Field (94.1%)	Two dimensional cepstral coefficient proved to be effective classifier feature for ESR

Continued on next page

Table 2.1 Literature review summary of audio event recognition.

Author	**Database**	**Features**	**Classifier**	**Accuracy**	**Remarks**
Cowling et al., [3]	8class, f_s = 44.1kHz, r=16 bit/sample	LTS, MFCC, FT, HCC, STFT, FWT, CWT	ANN, LVQ, GMM, LTS, DTW	MFCC with DTW 70%	DTW with MFCC is superior than ANN, LVQ, LTS and GMM for ESR
Bertelli et al., [4]	10class f_s = 8kHz, r=16 bit/sample	MFCC	FFNN	$73\% - 95\%$	Pattern recognition approach to identify environmental sound
Cakir et al., [5]	TUT-2009, 2016, 10 class, f_s = 22.05kHz	log mel band energy	FNN, CNN, RNN, CRNN	69.1%	CRNN give good performance for polyphonic sound
Souli et al., [6]	10class, f_s = 44.1kHz, r=16 bit/sample	MFCC, Spectrogram	SVM, HMM	MFCCs +SW (92.22%)	Application of scatter wavelet transform in audio enhancement framework
Wang et al., [7]	13class (audio scene, events and phone context)	Gaussian histogram	KNN	Audio scene (85.20%), events (77.6%), Phone context (88.9%)	Detection of surrounding environment like crowd

Continued on next page

Table 2.1 Literature review summary of audio event recognition.

Author	**Database**	**Features**	**Classifier**	**Accuracy**	**Remarks**
Valero et al., [8]	BBC Sound Effects Library, Freesound Project, $f_s = 11kHz$	MPEG-7, GTCC, MFCC	KNN, DT, SVM, NN	MFCC+ SVM (76.42%), GTCC+ SVM (78.01%)	Examine the superiority of GTCC compared with MFCC without increasing the computational cost
Lee et al., [9]	28class species, f_s = 44.1kHz, r= 16bit/sample	TDMFCC, DTDM-FCC	KNN	84.06%	Automatic Bird species classification
Ghoraani et al., [10]	10class, SAR (Ryerson University) f_s = 22.05kHz	Time-frequency features, MFCC	LDA	75.8%-85%	Address the trade-off between long-term analysis of sound, and their non-stationary features
Lin et al., [11]	Muscle fish, 16class, f_s = 8kHz, r=16 bit/sample	Subband power, Bandwidth, Pitch frequency, Brightness, FCC	Multiclass SVM	91%-97%	Wavelet with SVM accurately classify & categorize audio data accurately

Continued on next page

Table 2.1 Literature review summary of audio event recognition.

Author	**Database**	**Features**	**Classifier**	**Accuracy**	**Remarks**
Chu et al., [13]	14class f_s = 22.05kHz, r=16 bit/sample	MP features, TF dictionaries, MFCC	KNN, GMM	MP feature 72.5%, MFCC 70.9%	Multi audio category environment recognition with Matching Pursuit tool gives better performance than MFCC under some condition
Chu et al., [14]	14class, f_s = 22.05kHz, r= 16bit/sample	MFCC, MP and TF features,	KNN, GMM	66.80%-83.90%	Matching Pursuit technique with effective time frequency(TF) approach to achieve good accuracy for ESC
Wang et al., [16]	17class, f_s = 16kHz, r=16 bit/sample	Non uniform Scale-Frequency Map	Multiclass SVM	86.21%	Non uniform scale frequency map approach improve small accuracy
Wang et al., [17]	10class environmental sound, $f_s = 16kHz$	Wavelet Subspace-Based Features	Multiclass SVM	85.10% 90.63%	Robust system under noise for ESC

Continued on next page

Table 2.1 Literature review summary of audio event recognition.

Author	**Database**	**Features**	**Classifier**	**Accuracy**	**Remarks**
Rakotomamonjy, [18]	DCASE-16, Rouen-15, $f_s = 16kHz$	SNMF Matrix, HOG, PSD, RQA	CNN, SNMF	DCASE (SNMF-79.65% &CNN (79.16%), Rouen-15 (SNMF-78.59% &CNN (78.17%)	SNMF technique is less prone to over-fitting for smaller scale data-set than CNN model
Tran et al., [19]	10class environmental sound	Subband temporal envelop (STE), MFCC	SVM, GMM, polySVM	STE+SVM (75.8%), MFCC+ SVM (74.2%), STE+ polySVM (76.2%), MFCC+ GMM (75.9%)	SVM with STE feature perform well than SVM with MFCC
Mc Loughlin et al., [20]	RWCP, NOISEX-92	SAI, SIF	SVM , DNN	SAI+SVM (94.33%), SAI+DNN (96.20%), SIF+DNN (98.87%)	Comparison of auditory image and spectrogram image based end feature

Continued on next page

Table 2.1 Literature review summary of audio event recognition.

Author	**Database**	**Features**	**Classifier**	**Accuracy**	**Remarks**
Yang et al., [21]	TUT-16, 15class, $f_s = 44.1kHz$, r=24 bit/sample	MFCC, LBP, RQA	SVM	80.30%	Local binary pattern with temporal feature for ASC improves accuracy
Rabaoui et al., [22]	9class, NOISEX-92, $f_s = 44.1kHz$, r=16 bit/sample	PLPC, LPCC, MFCC, DWC ,MFDWC	SVM, HMM, MC-SVM	MC-SVM (96.89%)	MC-SVM multiclass outperform than binary class SVM and HMM in ESC
Sigtia et al., [23]	Baby cry, smoke alarm ,$f_s = 16kHz$	MFCC, ZCR, Statistical spectral moments	DNN, GMM, SVM, RNN	EE Ratio is DNN (10.8), SVM (12.9), GMM (14)	Investi gating the cost computation performance of the various classifier
Genco glue et al., [24]	61 class distinct sound	MFCC, Mel-Energy. log Mel-Energy	DNN	64.60%	DNN classifier for audio event detection outperform than GMM and HMM model
Souli et al., [25]	10class, $f_s = 44.1kHz$, r=16 bits/sample	Spectr ogram	MC- SVM	89.62%	Environment sound classification based on visual perception of spectrogram

Continued on next page

Table 2.1 Literature review summary of audio event recognition.

Author	**Database**	**Features**	**Classifier**	**Accuracy**	**Remarks**
Zhang et al., [26]	RWCP, NOISEX-92, $f_s = 16kHz$	Spectr ogram	CNN	93%	CNN with spectrogram feature demonstrate excellent performance under noise
Boddapati et al., [27]	ESC-50, ESC-10, Urbansound-8K, $f_s = 32kHz$, image $size = 256 \times 256$	Spectogram, MFCC & CRP	AlexNet & GoogLeNet	ESC-10(91%), ESC-50(73%), Urban sound-8K(93%)	Use of CNN classifier for both object and sound recognition

Table 2.2: Literature survey of standard environmental audio database.

Database	**Specifications**				
	Class	**Duration**	**SR(kHz)**	**BD(b/s)**	**Class Names**
AudioSet	632	4971 h			human, animals, natural, music, channel, environment and background, source-ambiguous sounds, sounds of things
Muscle Fish	16	100 min	8	16	altotrombone, animals, bells, cello bowed, crowds, female speech, laughter machine sounds , male speech, oboe, telephone, percussion, tubular bells, violin bowed, violin puzz, water
RWCP	14		16	16	bells, bottle, buzzer, cymbals, horn, kara, metal, phone, ring, whistle
NOISEX-92	8		16		speech noise, machine gun, STITEL , lynx, F16 , car, factory, operations room
TUT Acoustic Scenes , 2016	15	13 h	44.1	24	lakeside beach, bus, cafe/restaurant, car, city center, forest path, grocery store, home, library, metro station, office, urban park, residential area, train, tram

Continued on next page

Table 2.2 Literature survey of standard environmental audio database.

Database	**Specifications**				
	Class	**Duration**	**SR(kHz)**	**BD(b/s)**	**Class Names**
UrbanSound -8K	10	8.75 h	32	32	air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, street_music
NIGENS	14	4h : 45min	44.1	32	alarm, crying baby, crash, barking dog, running engine, burning fire, footsteps, knocking on door, female and male speech, female and male scream, ringing phone, piano
LITIS ROUE NORE	19	1500 min	22.05	24	plane, busy street, bus, cafe, car, train station hall, kid game hall, market, metro-paris, metro-rouen, billiard pool hall, quiet street, student hall, restaurant, pedestrian street, shop, train , high-speed train, tubestation
ESC-US	unlabeled	4166 min	48	32	
GTZAN	10	8.33 h	22.05	24	Music Genre
Magna TagATune	188	215 h			Music
Ballroom	8	5.8 h			Music

Continued on next page

Table 2.2 Literature survey of standard environmental audio database.

Database	**Specifications**				
	Class	**Duration**	**SR(kHz)**	**BD(b/s)**	**Class Names**
ESC-50	50	170 min	44.1	32	dog, rain, crying baby, door knock, helicopter, rooster, sea waves, sneezing, mouse click, chainsaw, pig crackling, fire, clapping, keyboard typing, siren, cow, crickets, breathing door, wood creaks, car horn, frog, chirping birds, coughing, can opening, engine, cat, water drops, footsteps, washing machine, train, hen, wind, laughing, vacuum cleaner, church bells, insects (flying), pouring water, brushing teeth, clock alarm, airplane, sheep, toilet flush, snoring, clock tick, fireworks, crow, thunderstorm, drinkings sipping, glass breaking, hand saw

Continued on next page

Table 2.2 Literature survey of standard environmental audio database.

Database	**Specifications**				
	Class	**Duration**	**SR(kHz)**	**BD(b/s)**	**Class Names**
TUT Acoustic Scenes 2017	15	52 min	44.1	24	bus, cafe/restaurant, car, city center, forest path, grocery store, home, lakeside beach, library, metro station, office, residential area, train, tram, and urban park
ESC-10	10	34 min	44.1	32	sneezing, dog barking, clock ticking, crying, rain, sea waves, fire crackling, helicopter, chainsaw baby, crowing rooster
The Duke noise DB	10				aircraft cockpit , babble , city rain, flat communications channel, helicopter fly, automobile highway, large city, large crowd, IBM PS/2 cooling fan, sun cooling fan, white Gaussian noise

Chapter 3

Audio Event Recognition Schemes

In this chapter, we presents audio event recognition (AER) schemes using the MFCC and spectrogram as audio features with machine learning and deep learning based classifiers. In this study, our objective is to develop effective and efficient audio event recognition scheme for automatically recognizing seven sound classes including *aircraft, construction, music, nature (wind and rain), speech, vehicle, and train* by choosing the suitable audio frame size. Since each of the sources can produce sounds with different duration which need to be considered in the creation of sound models for achieving better recognition accuracy with processing time. In some of the sound recognition methods like speech processing, selection of optimal speech frame size was well studied based on the glottal periods for processing and analysis of speech signals. But it is very difficult to have a complete study of the duration of each of the sounds that can be produced by different sources in real-life. Thus, we study the performance of the AER schemes under audio frame sizes of 100 ms, 250 ms and 500 ms. The each technique is discussed in detail in the next sections. The block diagram of the machine learning (ML) and deep learning (DL) classifiers based audio event recognition scheme is shown in Fig. 3.1, which consists of four major steps: pre-processing, feature extraction, sound models and recognition. Each of the audio event recognition steps is described in the next section.

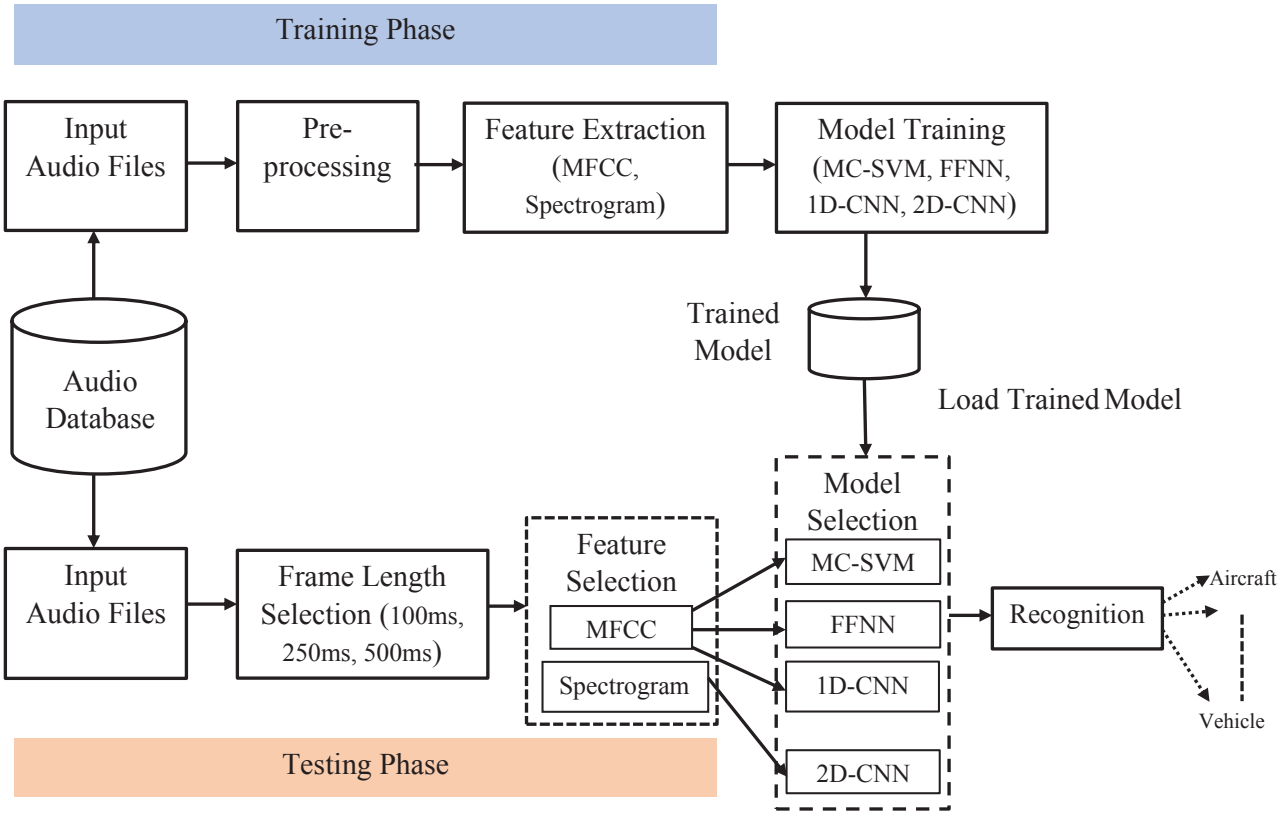


Figure 3.1: Block diagram of the audio event recognition schemes using the MFCC and spectrogram feature with machine learning and deep learning based classifiers.

3.1 Audio Features

Audio features should offer good representation of the audio signal. There are several steps to extract the features from a sound recording.

3.1.1 Preprocessing

Since an acoustic sensor is exposed to severe environmental sound sources like wind, the recorded audio signals may be corrupted with the low-frequency components having frequency below 10 Hz which are produced by the senors' movements due to the winds. Since the preservation of these low-frequency components may not be interested for our seven sound classification, the recorded audio signal is passed through high-pass filter with cut-off frequency of 10 Hz. Then, the recorded audio signal is divided into

frames with a predefined frame size (100/250/500 ms). In practice, sound intensity (or amplitude) depends on the location of an acoustic sensor from a sound source location which can be time-varying in the audio surveillance zone. Also the source location cannot be known apriori in realistic scenarios. Some times the low sound intensity can be recorded when an acoustic sensor is far away from a sound source although the sensitivity of a sensor is very good. In order to avoid the microphone sensitivity variation, the amplitude normalization is performed on the zero-mean audio signal.

3.1.2 Mel Frequency Cepstrum Coefficients Feature

Since our objective to investigate the performance of three ML classifiers for different frame size, we used the standard mel-frequency cepstral coefficients (MFCC) feature which is most commonly used in many environmental, speech and music sounds classification because human perception sensitivity with respect to frequencies is considered [5]. The basic structure of MFCC is shown in Fig. 3.2.

Each step of the MFCC feature extraction is described below:

- **Pre-emphasis filter** is used to amplify the high frequencies that balances the spectrum since high frequencies usually have smaller amplitudes as compared to the lower frequencies and may improves the signal-to-noise ratio (SNR). The pre-emphasis filter is implemented as [58]

$$y[n] = x[n] - \alpha x[n-1], \tag{3.1}$$

  where α is fixed as 0.97.

- After pre-emphasis, window function such as the Hamming window is applied to each audio frame. The Hamming window function is defined as [3]

$$h(n) = 0.54 - 0.46 \ cos\left(\frac{2\pi n}{N-1}\right) \tag{3.2}$$

where, $0 \leq n \leq N-1$, where N represents the length of audio frame. The window function $h[n]$ is multiplied with the filtered signal $y[n]$.

- **Fourier Spectrum**: The magnitude spectrum is computed by taking the fast Fourier transform (FFT) of the windowed audio frame that corresponds to different energy distribution over frequencies.

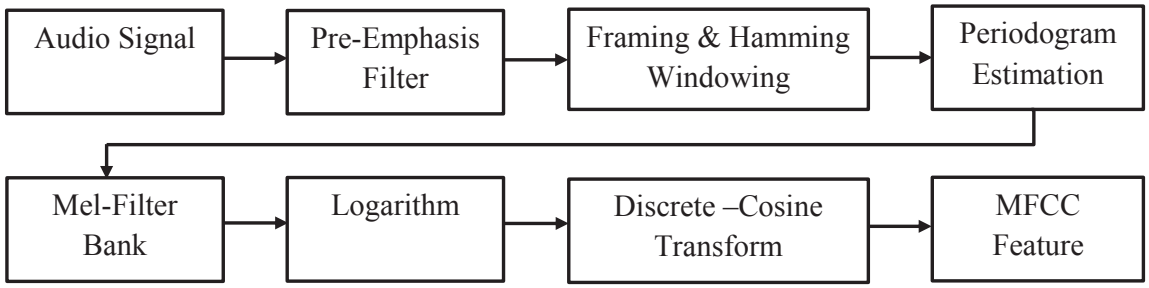


Figure 3.2: Basic block diagram of Mel-Frequency-Cepstrum- Coefficient. This figure is adapted from [6].

- **Mel-Frequency Spectrum**: The magnitude spectrum is multiplied by a set of 26 triangular band-pass filters which have the positions equally spaced along the Mel-scale which is related to the linear frequency f by the following equation [19]

$$M(f) = 1125 \ \ln\left(1 + \frac{f}{700}\right). \tag{3.3}$$

The Mel-frequency is proportional to the logarithm of the linear frequency that reflects the similar audio effects in the human's subjective perception.

- **MFCC**: Since the filter bank coefficients are highly correlated the discrete cosine transform (DCT) is applied to decorrelate the filter bank coefficients. The coefficients are computed as [5]

$$C_n = \sum_{k=1}^{K} (\log \ S_k) \left[n \left(k - \frac{1}{2} \right) \frac{\pi}{K} \right] \tag{3.4}$$

where, $n = 1, 2....., K$ and $K = 26$, number of triangular band-pass filters, S_k energy output of the K^{th} triangular band-pass filter.

Thus, above feature extraction step results in 26 coefficients. In this study, we use only

the lower 13 of the 26 coefficients for each frame. The parameters used for extracting the MFCC are: pre-emphasis value of 0.97, 26 filters, lower band edge of 0 Hz, higher band edge of 8000 Hz, and 13 cepstral coefficients for each audio frame. The MFCC plot of each audio event shown in Fig. 3.3.

3.1.3 Spectrogram Image Feature (SIF)

A Spectrogram is a visual representation of the intensity in the spectrum of frequencies, of a sound, that varies with time [44]. Spectrogram is generated from an audio signal using Short Time Fourier Transform (STFT) [26]. In forming the spectrogram image, discrete Fourier transform (DFT) is applied to the windowed signal as [43]:

$$X(k,t) = \sum_{n=0}^{N-1} x(n)w(n)exp(-\frac{j2\pi kn}{N}) \tag{3.5}$$

where hanning window function w(n) is defined as [25]:

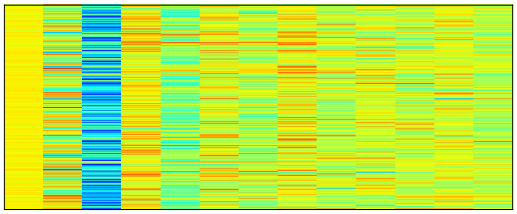
$$w(n) = 0.5 - 0.5\ cos\left(\frac{2\pi n}{N-1}\right) \tag{3.6}$$

$0 \leq n \leq N-1$, where N is the length of the window, x(n) is the time-domain signal, X(k,t) is the k^{th} harmonic corresponding to the frequency $f(k) = kf_s/N$ for the t^{th} frame, f_s is the sampling rate.

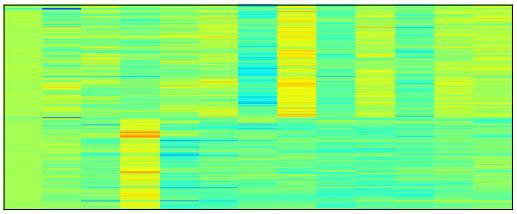
The STFT of an acoustic event is computed using Hanning window of size 100ms with 50% overlap and 16000 Hz sampling rate. This gives the spectrum of complex values (X(k, t)). The magnitude of STFT yields linear spectrogram and log of linear spectrogram yields logarithmic spectrogram S(k, t) as [44]:

$$S(k,t) = log(|X(k,t)|) \tag{3.7}$$

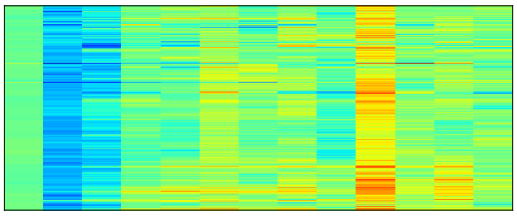
where k is frequency bin and t is the time frame. The log operation reduces the dynamic range of spectrogram energies and enhances the spectral components belonging to an



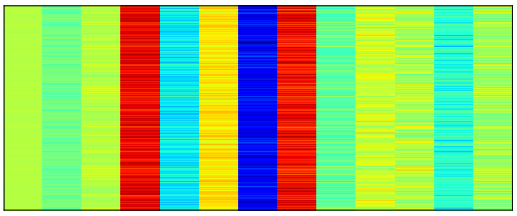
(a) Aircraft Sound



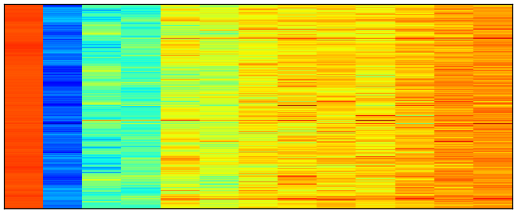
(b) Train Sound



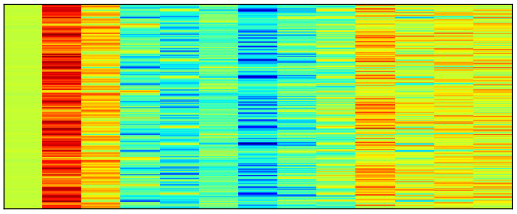
(c) Bus Sound



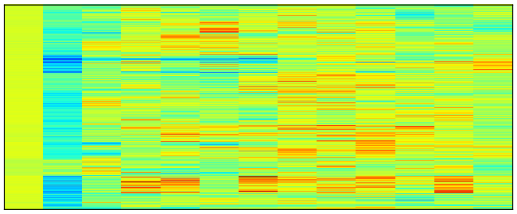
(d) Tractor Sound



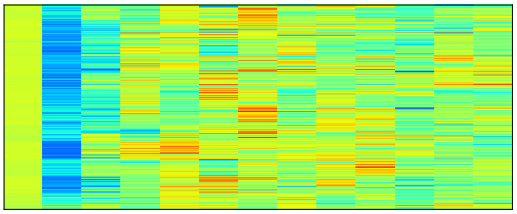
(e) Rain Sound



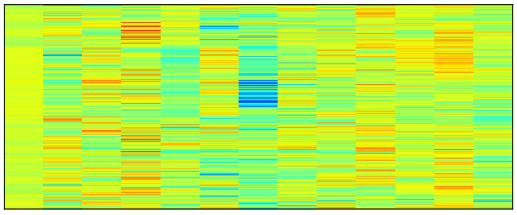
(f) Wind Sound



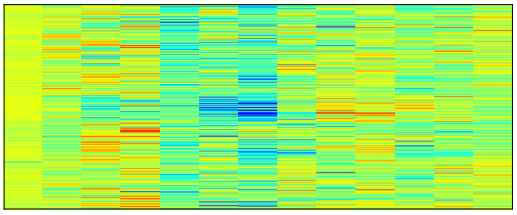
(g) HipHop Music Sound



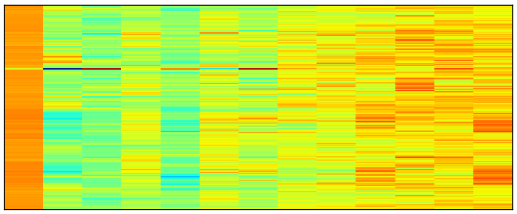
(h) Disco Music Sound



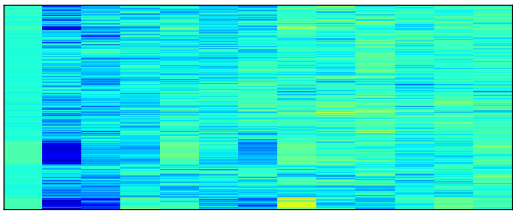
(i) Male Speech Sound



(j) Female Speech Sound



(k) Drilling Sound



(l) Hammer Sound

Figure 3.3: MFCC plot of audio events.

acoustic event. The same procedure is followed for the generation of spectrogram image for 250 ms and 500 ms frame length. The spectrogram image is obtained with dimension 128×128 using above procedure. The spectrogram image of audio events are shown in Fig. 3.4.

3.2 Machine Learning Classifier

In this study, we evaluated the performance of the machine learning based classifiers such as multi-class support vectors machines (MC-SVM) under different audio frame size of 100, 250 and 500 ms for automatically recognizing the 7 sound classes using MFCCs audio feature .

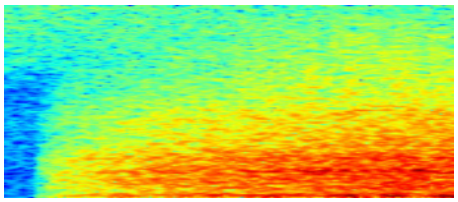
3.2.1 Multi-Class Support Vector Machine

Support vector machine is a type of supervised machine learning classification algorithm, introduced in 1960s [86]. In literature, SVM has been successfully used in pattern recognition applications, such as speaker identification, object detection, speech recognition, and text classification [87]. For a given training data (x_i, y_i) for $i = 1, 2, ...N$, the optimization problem for the SVM is formulated as follows [89]:

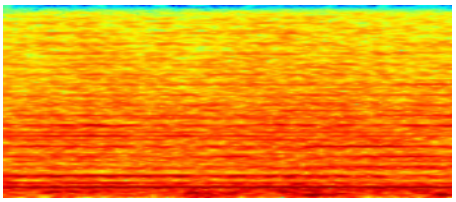
$$\arg\min_{w,\xi,b}\{\frac{1}{2}\|\mathrm{w}\|^2 + \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{N}}\xi_{\mathrm{i}}\} \tag{3.8}$$

subject to the constraints

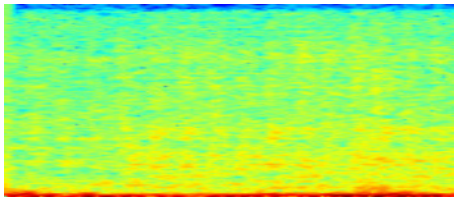
$$y_i(w \cdot x_i - b) \geq 1 - \xi_i, \qquad \xi_i \geq 0 \tag{3.9}$$



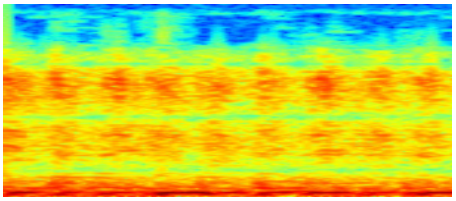
(a) Aircraft Sound



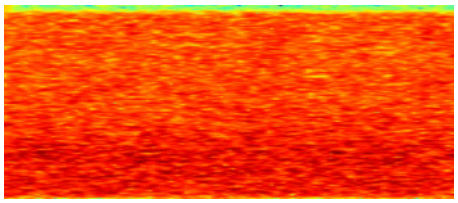
(b) Train Sound



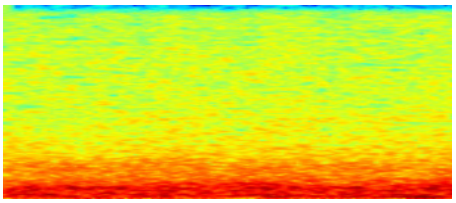
(c) Bus Sound



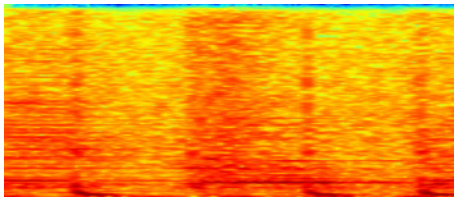
(d) Tractor Sound



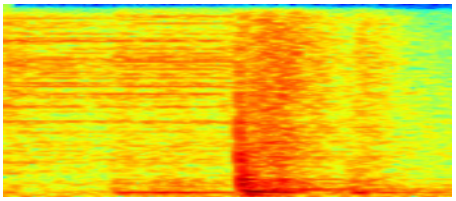
(e) Rain Sound



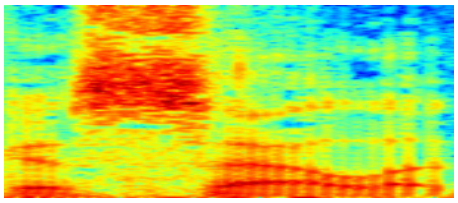
(f) Wind Sound



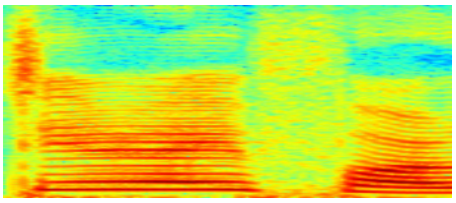
(g) HipHop Music Sound



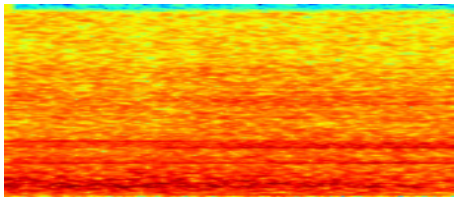
(h) Disco Music Sound



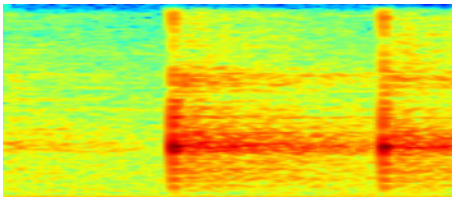
(i) Male Speech Sound



(j) Female Speech Sound



(k) Drilling Sound



(l) Hammer Sound

Figure 3.4: Spectrogram of each audio events.

where ξ_i are non-negative slack variables.It is solved using a Lagrangian formulation of the problem, thus producing the multipliers α_i and the decision function [91]:

$$f(x) = sgn\left(\sum_{i=0}^{N} y_i \alpha_i x \cdot x_i + b\right) \tag{3.10}$$

where N is the number of training samples and x is a feature vector. A nonlinear kernel function $\mathrm{K}(x_i, x_j)$ is used to replace the dot products $x \cdot x_i$, with the effect of projecting the data into a higher dimensional space where it is linearly separable. The decision function is defined as [90]:

$$f(x) = sgn\left(\sum_{i=0}^{N-1} y_i \alpha_i K(x, x_i) + b\right) \tag{3.11}$$

In this thesis work, the Gaussian radial basis function is used as kernel. It is defined as $K_R(x, x_i) = exp(-\gamma \|x - x_i\|^2)$, and the MC-SVM classifier was implemented using Scikit-learn, which is one-against-rest multi-class classifier [59] , [88]. For the creation of MC-SVM model whole dataset is divided into development and evaluation dataset. The model is trained with 28 hours of audio data and 5-fold cross validation is performed on the development data, i.e., 5 different combinations are made on the selection of data so that each portion is used in a testing dataset at least once. First fold is used to tune the parameters and the network is evaluated for the other four folds. The final results are presented by taking the average of the results for the last four folds as shown in Fig. 3.5.

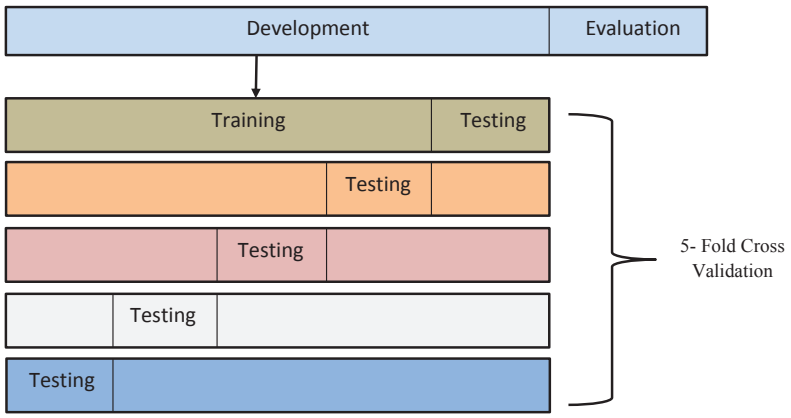


Figure 3.5: Database partitioning into development and evaluation set in 5-fold training. This figure is adapted from [52]

The hyperparameter which is used during training of MC-SVM model are penalty parameter 1, gamma 0.0769, degree 3, tolerance 0.001 with infinite epochs. The MC-SVM model based on 500 ms frame size take less time for training compare to other MC-SVM model. The specifications of the MC-SVM based AER scheme are summarized in Table 3.1.

Table 3.1: Specifications of Multi-class SVM based AER Scheme

Parameters	**Frame Length**		
	100ms	**250ms**	**500 ms**
Audio format	wav	wav	wav
Channel	Mono	Mono	Mono
Bit depth	16 bit PCM	16 bit PCM	16 bit PCM
Class	7	7	7
Events	28	28	28
Training and validation data duration	28 h	28 h	28 h
Testing data duration	70 min	70 min	70 min
Sampling rate	16 kHz	16 kHz	16 kHz
Feature	MFCC	MFCC	MFCC
Number of samples in a frame	1600	4000	8000
NFFT	2048	4096	8192
Kernel	RBF	RBF	RBF
Penalty parameter C	1	1	1
Gamma	0.0769	0.0769	0.0769
Degree	3	3	3
Fold	5	5	5
Cache size	200	200	200
Tolerance	0.001	0.001	0.001
Epochs	Infinite	Infinite	Infinite
Training time	196 h	29 h	10 h
Trained model size	60.1 MB	25.3 MB	14.3 MB

3.3 Deep Learning Classifiers

Deep learning is the advance version of machine learning [84]. Deep Learning is about learning multiple levels of representations and abstractions that help to make sense of data. Many layers are used that compute non-linear functions and model highly complex data. Essentially, deep neural networks have been inspired from the human brain architecture [67]. In deep neural networks number of hidden layers are more compare to artificial neural networks (ANNs). ANNs contain single hidden layer is the universal approximator. But sometimes it can't handle the feature of large dataset. The goal of introducing multiple hidden layers is to find more abstract features in the higher levels. To perform same as DNNs, ANNs require vast amount of neurons to represent of the input. Due to increase in number of neurons , the parameters of network significantly increases which increase the computational complexity. In literature it is discussed that ANNs with vast amount of neurons not efficient. The enhancement in computational technology paved the way for deep learning architectures. With the advancement of parallel computing technologies, the limitation of conventional ANN is relatively solved. Nowadays we can use hidden layers upto 150 [51] .

3.3.1 Feed Forward Neural Network Based AER Scheme

The FFNN architecture contains the input layer, hidden layer and output layer. Each layer gets its input from a previous layer and then computes and transforms the data. The transformed data sends to the next layer. Each layer consists of neurons which are having various modes of connections to other neurons in the same layer as well as the other layers depending on the type of network. Each layer connection has its own weights. The block diagram of feed forward neural network used in this work is shown in Fig. 3.6.

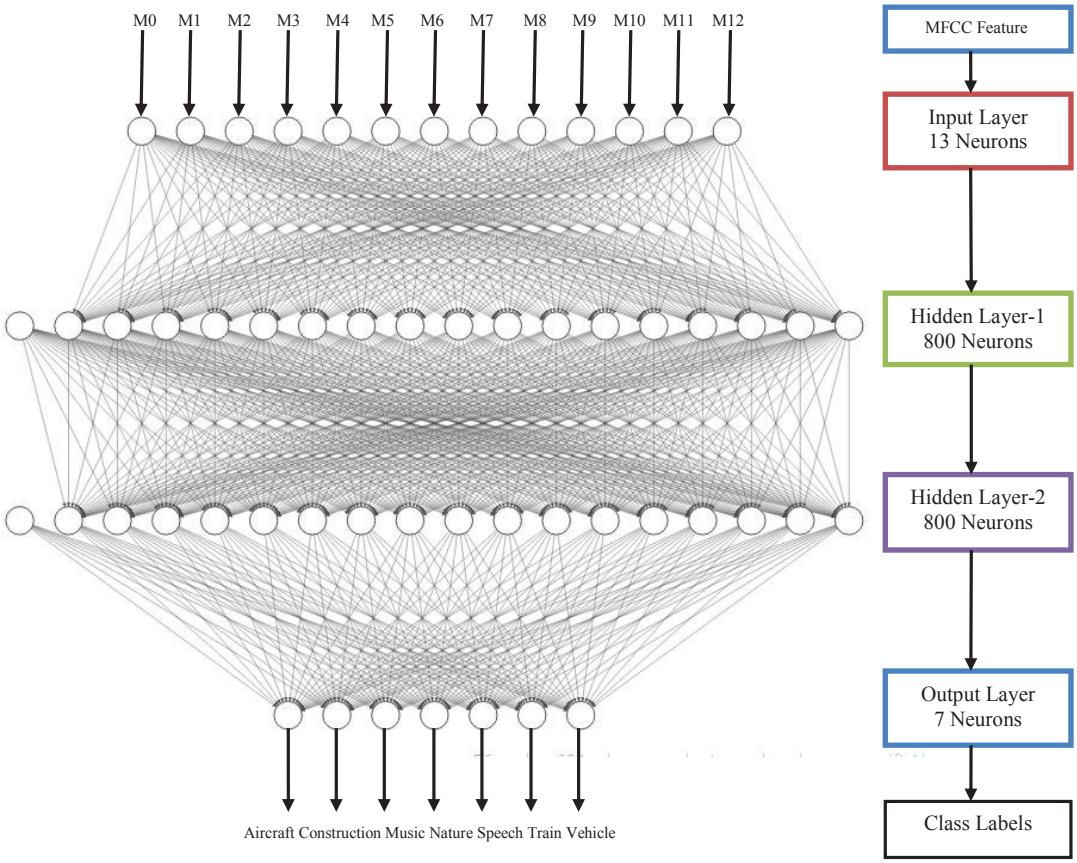


Figure 3.6: Block diagram of the FFNN based AER scheme.

The non-linear transformation is defined as [4]:

$$h^i = f(W^i h^{i-1} + b^i), \quad for \ 0 \leq i \leq L \tag{3.12}$$

where h^0 corresponds to the input x, W_i, b_i are the weight matrix and bias vector for the i^{th} layer, and the output of the final layer h^L is the desired output. The non-linearity f is usually a *sigmoid* or a *hyperbolic tangent function, tanh(x)*. However, the Rectified Linear Unit (ReLU) function is used for faster convergence of models in the training phase. The $ReLU$ function is defined as $f(x) = max(0, x)$, which is simpler than the *sigmoid or tanh activation* because it only requires a piece-wise linear operator instead of a fixed point *sigmoid or tanh*. Feed forward neural network also known as fully connected neural network or multilayer perceptron network (MLP).

In this work, the FFNN classifier was implemented using Keras library in backend of Tensorflow framework. For the creation of FFNN model whole dataset is divided

into development and evaluation datasets. The model is trained with 70 hours of audio data and 5-fold cross validation is performed on the development data as shown in Fig. 3.5. We fixed our FFNN model with two hidden layers with each layer contains 800 neurons. ReLU activation function were applied on top of each hidden layer. The output layer is softmax layer with categorical cross entropy loss function that outputs the subject ID ranging from 0 to 99. We used the Adam optimizer, which popular in deep learning framework, learning rate 0.001, batch size 128 with 15 epochs. We also set the early stopping while training the model to avoid the overfitting. The FFNN model learning curve is shown in Fig. 4.3, 4.4 , and 4.5. The specifications of the FFNN based AER scheme are summarized in Table 3.2.

Table 3.2: Specifications of the FFNN based AER scheme

Parameters	**Frame Length**		
	100ms	**250ms**	**500 ms**
Audio format	wav	wav	wav
Channel	Mono	Mono	Mono
Bit depth	16 bit PCM	16 bit PCM	16 bit PCM
Class	7	7	7
Events	28	28	28
Training and validation data duration	70 h	70 h	70 h
Testing data duration	70 min	70 min	70 min
Sampling rate	16 kHz	16 kHz	16 kHz
Feature	MFCC	MFCC	MFCC
Number of samples in a frame	1600	4000	8000
NFFT	2048	4096	8192
Number of hidden layer	2	2	2
Fold	5	5	5
Activation functions	Softmax, Relu	Softmax, Relu	Softmax, Relu
Number of neurons	1600	1600	1600
Batch size	128	128	128
Learning rate	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam
Epochs	15	15	15
Number of FFNN parameters	657607	657607	657607
Training time	277 min	72 min	57 min
Trained model size	7.92 MB	7.92 MB	7.92 MB

3.3.2 Convolutional Neural Networks Based AER Scheme

The convolutional neural networks (CNNs) technique is made up of two components. The first component is the feature identifier where the features from the input data are automatically learned [47]. The second component is a fully connected multi-layer perceptron (MLP) which carry out classification based on the initially learned features. Further, the feature identifier component comprises of convolutional and pooling layers. In the convolutional layer, the activation (or feature) map from the previous layer is convolved using convolutional filter (or kernel) which is added with a bias and subsequently fed to the activation function to generate an activation map for the next layer [48]. Meanwhile, the pooling layer (or subsampling layer) causes the activation maps to be reduced but, increases the invariance to distortion in the inputs. The convolutional and pooling layers are positioned to accomplish high level feature extraction [43]. Also, a simple classifier, such as softmax, can be used in the last part of the CNNs.

Let $x_i^0 = (x_1, x_2....., x_n)$ be the data input vector, where n is the total number of samples in the audio segment. Next, the convolutional layer output is computed as follows [26],

$$c_i^{l,k} = \sigma(b_k + \sum_{n=1}^{N} w_n^K X_{i+n-1}^{0k}) \tag{3.13}$$

where l is the layer index, σ is the activation function producing non-linearity, b is the bias for the k^{th} activation map, N is the size of the filter, w_n^k is the weight for the nth filter index and kth activation map. Also, max pooling can be used to compute the maxi- mum value in an input. The activation map in a layer is the pool using the following computation [34],

$$P_i^{l,k} = \max_{t \in T}\{ \mathrm{c}_{\mathrm{iXS+r}}^{\mathrm{l,k}}\} \tag{3.14}$$

Where T is the window size of the pooling and S is the pooling stride. Thus, the

activation map from layer to layer forward propagation is computed using above equation [37]. This includes initializing the weights and calculating the error cost minimization by using stochastic gradient descent on the audio event [20]. After obtaining the predicted output, the loss function is used to calculate the prediction error [26]. Then, back propagation is implemented to adjust the weights and the error is predicted by calculating the slope of the convolutional weights [5]. The process of forward and back propagation is continuously executed till the required number of epochs or other stopping criteria is met [24].

3.3.3 One-dimensional Convolutional Neural Network

Fig. 3.7 shows a block diagram of the one-dimensional CNN based AER scheme. The details of the 1D-CNN based AER scheme are summarized in Table 3.3. In this study, the number of layers and the tuning parameters are varied for the optimum performance. Hence, the proposed model consists of 4 convolutions, 1 max-pooling, 2 dropout, 1 flatten and 2 fully-connected layers. The filter shifts are set at 1 and 2 for convolution and max-pooling, respectively. The max-pooling operation reduces the dimensions of feature maps and also retains the important and significant features of the audio feature. The flatten layers convert 2D output into 1D output. To avoid the the model overfitting, a dropout layer is used before each dense layer. In the last stage, the fully-connected layer is used to connect the neurons in the previous layers.

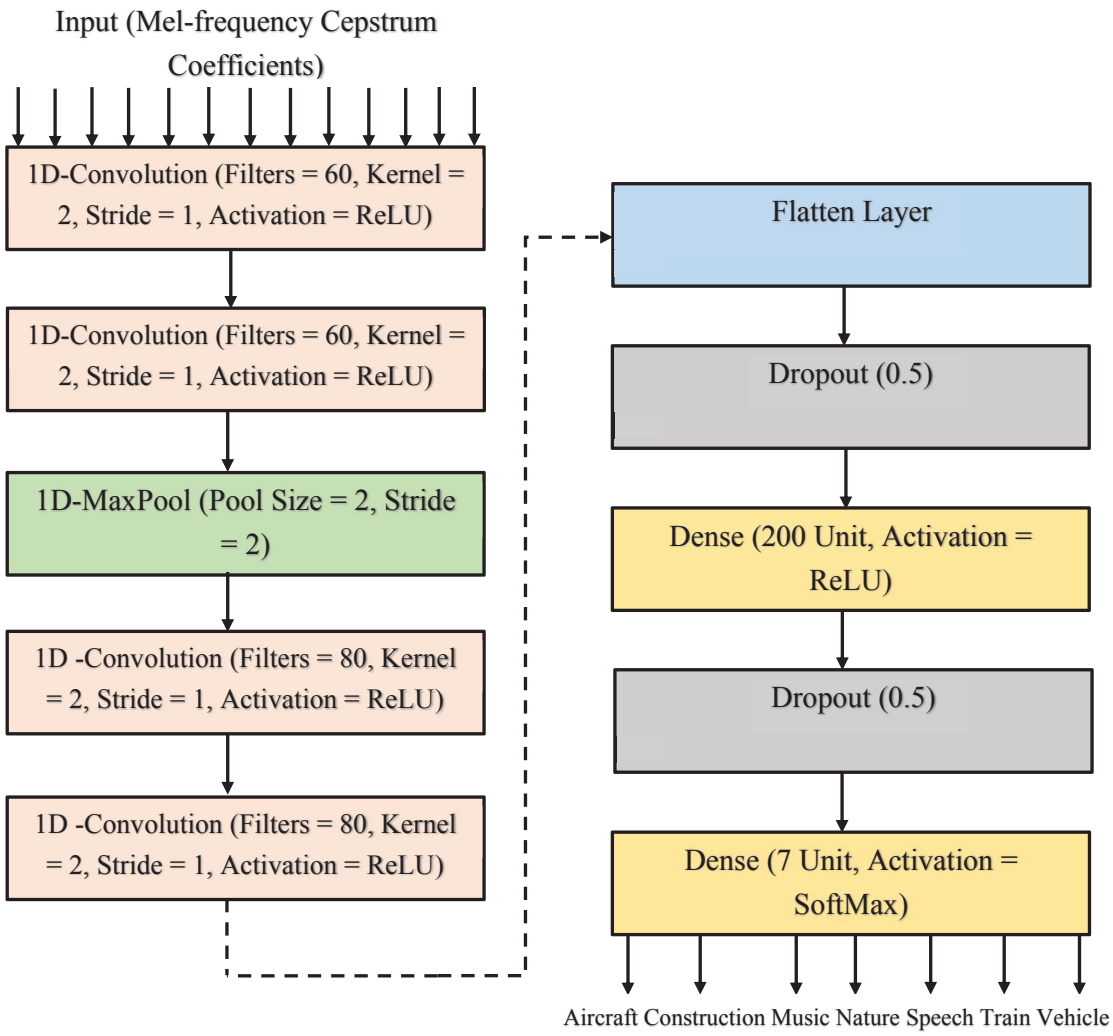


Figure 3.7: One-dimensional CNN based audio event recognition (AER) scheme.

In this work, the 1D-CNN classifier was implemented using Keras library in backend of Tensorflow framework. For the creation of 1D-CNN model whole dataset is divided into development and evaluation datasets. The model is trained with 70 hours of audio data and 5-fold cross validation is performed on the development data as shown in Fig. 3.5. For creating the model, we consider a local filter with kernel size 2 across across all MFCC coefficients, which is 5039930/13 in our dataset. We used our model with 4 1D- convolution layers with first two individual layer contained 60 feature maps and next two 80 feature maps. 1D-Max-pooling and ReLU modules are applied after first two convolution layer. Flatten layer is used after last convolution layer. The top layer contained 1 dense layer (FC1) with 200 hidden units. ReLU modules were also applied on dense layer. Dropout layer are only applied before the dense and output layer. The output layer is softmax layer with categorical cross entropy loss function that outputs

the subject ID ranging from 0 to 99. We used the Adam optimizer, commonly used in image recognition, and 0.001 for learning rate with 20 epochs, and the batch size is 128. We also set the early stopping while training the model to avoid the overfitting. The 1D-CNN model learning curve for different frame length is shown in Fig. 4.7 , 4.8 and 4.6. The specifications of the 1D-CNN based AER scheme are summarized in Table 3.4.

Table 3.3: Specification of 1D-CNN based AER scheme for different lengths

No.	**Layer (Type)**	**Frame Length**					
		100 ms		**250 ms**		**500 ms**	
		Shape	**Param**	**Shape**	**Param**	**Shape**	**Param**
1	Input	(13 ,1)	0	(13 , 1)	0	(13 , 1)	0
2	Conv1D	(12 , 60)	180	(12 ,60)	180	(12 , 60)	180
3	Conv1D	(11, 60)	7260	(11 , 60)	7260	(11 , 60)	7260
4	MaxPooling	(5 , 60)	0	(5 , 60)	0	(5 , 60)	0
5	Conv1D	(4 , 80)	9680	(4 , 80)	9680	(4 , 80)	9680
6	Conv1D	(3 , 80)	12880	(3 , 80)	12880	(3, 80)	12880
7	Flatten	240	0	240	0	240	0
8	Dropout	240	0	240	0	240	0
9	Dense	200	48200	200	48200	200	48200
10	Dropout	200	0	200	0	200	0
11	Dense	7	1407	7	1407	7	1407

Table 3.4: The summary of database and hyper-parameters of 1D-Convolutional Neural Network model used in this work

Parameters	**Frame Length**		
	100ms	**250ms**	**500 ms**
Audio format	wav	wav	wav
Channel	Mono	Mono	Mono
Bit depth	16 bit PCM	16 bit PCM	16 bit PCM
Class	7	7	7
Events	28	28	28
Training and validation data duration	70 h	70 h	70 h
Testing data duration	70 min	70 min	70 min
Sampling rate	16 kHz	16 kHz	16 kHz
Feature	MFCC	MFCC	MFCC
Number of samples in a frame	1600	4000	8000
NFFT	2048	4096	8192
Convolution layer	4	4	4
Dropout layer	2	2	2
MaxPooling layer	1	1	1
Fold	5	5	5
Activation functions	Softmax, Relu	Softmax, Relu	Softmax, Relu
Number of neurons	1600	1600	1600
Batch size	128	128	128
Learning rate	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam
Epochs	20	20	20
Number of FFNN parameters	79607	79607	79607
Training time	13 hours	2 hour	1 hour
Trained model size	1 MB	1 MB	1 MB

3.3.4 Two-dimensional Convolution Neural Network

In this work to recognized audio event using spectrogram image as feature we used 2D-CNN model with 14 layers the graphical representation of the architecture can be seen in Figure 3.8 and the details of the proposed CNN model are tabulated in Table 3.5. The number of layers and the tuning parameters are varied by a brute force method until the optimum diagnostic performance is achieved. Hence, the proposed model consists of 4 convolutions, 2 max-pooling, 4 dropout, 1 flatten and 3 fully-connected layers. These layers make up the fundamental structure of CNN whereby convolution picks up distinctive features from the input audio feature spectrogram. The max-pooling operation reduces the dimensions of feature maps and at the same time retain important and significant features of the input audio feature spectrogram. Flatten layers turns the 2D output into 1D outputs which allows us to add one or more dense layers to the model for classification. To avoid the the overfitting in CNN model one dropout layer is used before each convolution layer and output layer. Lastly, the fully-connected layer is intended to connect the neurons in the previous layers into a seven-class probability distribution. The specifications of the 2D-CNN based AER scheme are summarized in Table 3.6.

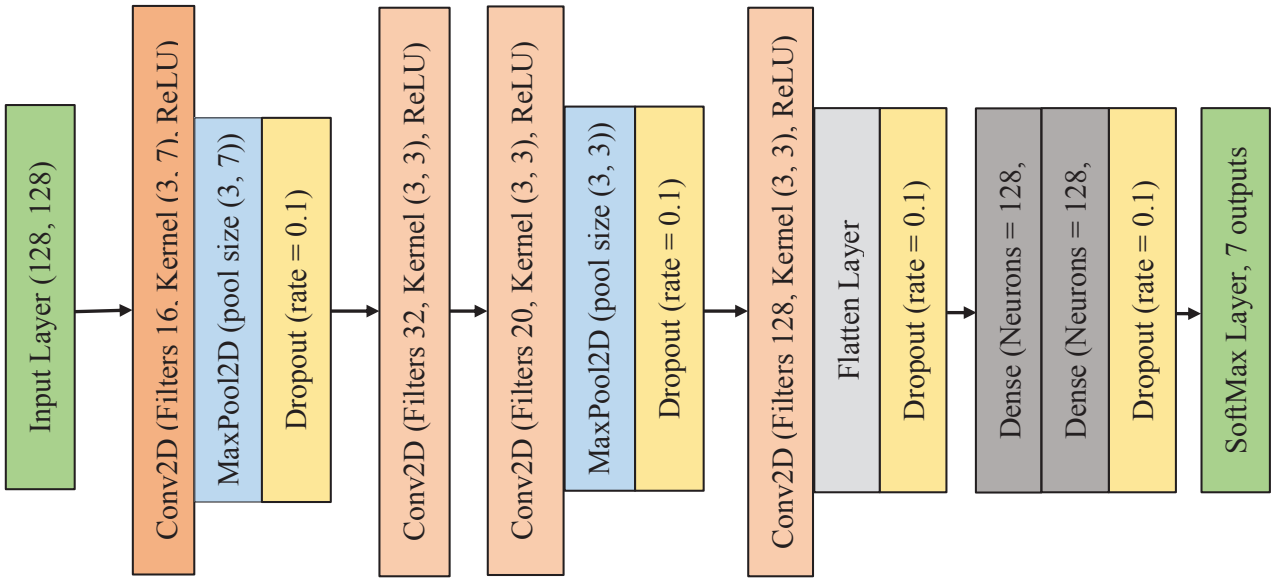


Figure 3.8: Two-dimensional CNN based audio event recognition (AER) scheme.

In this thesis, the 2D-CNN classifier was implemented using Keras library in backend of Tensorflow framework. For the creation of 2D-CNN model whole dataset is divided into development and evaluation datasets. The model is trained with 7 hours of audio data. For creating the model, we consider a local filter with kernel size (3,7) across across all spectrogram image, which is 128/128 in our dataset. We used our model with four 2D-convolution layers in which first layer contained 16 feature maps, second 32, third 20 and last 128 feature maps. 2D-Max-pooling and dropout layer are applied after first and third convolution layer respectively. Flatten layer is used after last convolution layer. The top layer contained 2 dense layer (FC1 and FC2) with 128 hidden units. ReLU modules were also applied on each dense layer. Dropout layer are only applied to the each dense layer. The output layer is softmax layer with categorical cross entropy loss function that outputs the subject ID ranging from 0 to 99. We used the Adam optimizer, commonly used in image recognition, and 0.001 for learning rate with 30 epochs, and the batch size is 21. The 2D-CNN model learning curve for different frame length is shown in Fig. 4.9, 4.10, and 4.11.

Table 3.5: Specification of 2D-CNN based AER scheme for different lengths

No.	**Layer(Type)**	**Frame Length**					
		100ms		**250ms**		**500ms**	
		Shape	**Param**	**Shape**	**Param**	**Shape**	**Param**
1	Conv2D	(126, 122, 16)	1024	(126, 122, 16)	1024	(126, 122, 16)	1024
2	MaxPooling2D	(42, 17, 16)	0	(42, 17, 16)	0	(42, 17, 16)	0
3	Dropout	(42, 17, 16)	0	(42, 17, 16)	0	(42, 17, 16)	0
4	Conv2D	(40, 15, 32)	4640	(40, 15, 32)	4640	(40, 15, 32)	4640
5	Conv2D	(38, 13, 20)	5780	(38, 13, 20)	5780	(38, 13, 20)	5780
6	MaxPooling2D	(12, 4, 20)	0	(12, 4, 20)	0	(12, 4, 20)	0
7	Dropout	(12, 4, 20)	0	(12, 4, 20)	0	(12, 4, 20)	0
8	Conv2D	(10, 2, 128)	23168	(10, 2, 128)	23168	(10, 2, 128)	23168
9	Flatten	2560	0	2560	0	2560	0
10	Dropout	2560	0	2560	0	2560	0
11	Dense	128	327808	128	327808	128	327808
12	Dense	128	16512	128	16512	128	16512
13	Dropout	128	0	128	0	128	0
14	Dense	7	903	7	903	7	903

Table 3.6: The summary of database and hyper-parameters of 2D-CNN model used in this work

Parameters	**Frame Length**		
	100ms	**250ms**	**500 ms**
Audio format	wav	wav	wav
Channel	Mono	Mono	Mono
Bit depth	16 bit PCM	16 bit PCM	16 bit PCM
Class	7	7	7
Events	28	28	28
Training and validation data duration	7 h	7 h	7 h
Testing data duration	70 min	70 min	70 min
Image width	128	128	128
Image height	128	128	128
Sampling rate	16 kHz	16 kHz	16 kHz
Feature	Spectrogram	Spectrogram	Spectrogram
Convolution layer	4	4	4
Dropout layer	4	4	4
MaxPooling layer	2	2	2
Activation functions	Softmax, Relu	Softmax, Relu	Softmax, Relu
Batch size	21	21	21
Learning rate	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam
Epochs	30	30	30
Number of 2D-CNN parameters	379835	379835	379835
Training time	144 h	20 h	5.6 h
Trained model size	4.62 MB	4.62 MB	4.62 MB

Chapter 4

Results and Discussions

In this chapter, we evaluate the performance of the machine learning classifier and deep learning classifiers based four AER schemes using a wide variety of audio recorded using different kinds of audio recording devices.

4.1 Experimental Setup

The specification of experimental setup used in this thesis work is given below:

Hardware: Desktop PC

RAM : 8 GB

OS Type : 64 bit

Processor : Intel Xeon(R) CPU E5-2620 v3 @2.4GHz ×12

Graphic Card : NVIDIA Quadro K-2200/PCIe/SSE2

Softwares:

OS : Ubuntu 16.04 LTS

Anaconda Python (Python 3.6.5, JupyterNotebook)

Deep Learning Framework : Tensorflow (1.12.0), Keras (2.2.0) [82]

Machine Learning Framework : Scikit-learn [59], [60]

Audio features library : `python_speech_feature` [58]

Audio analysis package : LibROSA [61]

Matlab 2015b

Audacity, WaveSurfer

4.2 Audio Database Collection

The description of the audio database is summarized in Table 4.1. Our audio signals are digitized at sampling rate of 44.1 kHz and 16-bit resolution. For training and testing purposes, the recorded audio signals are resampled to 16 kHz. We used H1n Handy recorder, EVISTR digital voice recorder and mobile handset (Samsung, Redmi) for recording audio under different environmental conditions. In addition with our audio signals, we also collected from the public multimedia websites (like YouTube, GTZAN library, freesound.org). A total duration of the audio is about 87 hours.

Table 4.1: Detail of collected audio database

Title	Class	Duration (h)	Contribution	
			Self	Internet
Aircraft	Aircraft	13		100%(YouTube)
Construction	Hammer, Concrete Mixer, Hand-saw, Drilling Machine, Hoe, Axe	10	100%	
Music	Blues, Clas-sic, Coun-try, Disco, Hiphop, Jazz, Metal, Pop, Reggae, Rock	10		100%(GTZAN)
Nature	Rain Wind, Thun-derstorm	12 2	70%	30%(YouTube)
Speech	Male Female	7.5 3.5	80%	20% (YouTube)
Vehicles	Bus Motorcycle Tractor JCB Auto-rickshaw Van	3 1 5 1 0.25 1	90%	10%(YouTube)
Train	Train	12	80%	20% (YouTube)
	Inverter AC	1 1		

4.3 Performance Evaluation

The performance of audio event recognition methods are evaluated by using benchmark metrics such as precision, recall (sensitivity), F1-score, and overall accuracy [52].

4.3.1 Confusion Matrix

A confusion matrix best represents the outputs of predictive analysis algorithms [84]. In this work to evaluate the benchmark metrics of multi-class problem the general confusion matrix used is shown in Fig. 4.1. It contains some specific terminology which are briefly explained below.

- True Positive (TP) is when the predicted output and the true output are both positive. In the Figure 4.1, the diagonal elements of confusion matrix are TP (i.e., TP_A (for Aircraft class), TP_C (for Construction class), etc.).
- False Positive (FP) is when the predicted output is positive while the true output is negative. The total number of FP for a class is sum of values in corresponding column except TP in that column. For example FP for the case of Aircraft (A) class is given as:

$$FP_A = ER_{CA} + ER_{MA} + ER_{NA} + ER_{SA} + ER_{TA} + ER_{VA} \tag{4.1}$$

- False Negative (FN) is when the predicted output is negative while the true output is positive. The total number of FN for a class is sum of values in corresponding row except TP in that row. For example FN for the case of Aircraft (A) class is given as:

$$FN_A = ER_{AC} + ER_{AM} + ER_{AN} + ER_{AS} + ER_{AT} + ER_{AV} \tag{4.2}$$

- True Negative (TN) is when the predicted output and the true output are both

negative. The total number of TN for a specific class is sum of all columns and rows excluding that class's column and row. For example TN for the case of Aircraft (A) class is given as:

$$TN_A = \text{Total Predictions} - (TP_A + FP_A + FN_A) \tag{4.3}$$

- The total number of test sample of any class would be sum of corresponding row (i.e., $TP + FN$ for that class)

The performance of a classifier is evaluated using certain performance parameters as discussed below :

1. Accuracy: It is the ratio between correctly predicted outcomes and sum of all predictions. It shows, overall how often is the classifier correct.

$$\text{overall accuracy (OA)} = \frac{\text{All TPs class}}{\text{Total matrix sum}} \tag{4.4}$$

2. Precision: It is defined as out of all the classes, how much model predicted correctly. It should be high as possible.

$$\text{precision (Pr)} = \frac{\text{TP}}{\text{TP + FP}} \tag{4.5}$$

3. Recall (Sensitivity): It is defined as out of all the positive classes, how much model predicted correctly. It should be high as possible.

$$\text{recall (Re)} = \text{sensitivity (Se)} = \frac{\text{TP}}{\text{TP + FN}} \tag{4.6}$$

4. F1-score: It is difficult to compare two models with low precision and high recall or vice-versa. So to make them comparable, F1-Score is used. F1-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place

of Arithmetic Mean by punishing the extreme values more.

$$\text{F1-score (F1)} = \frac{2\text{Re} \times Pr}{\text{Re} + \text{Pr}} \tag{4.7}$$

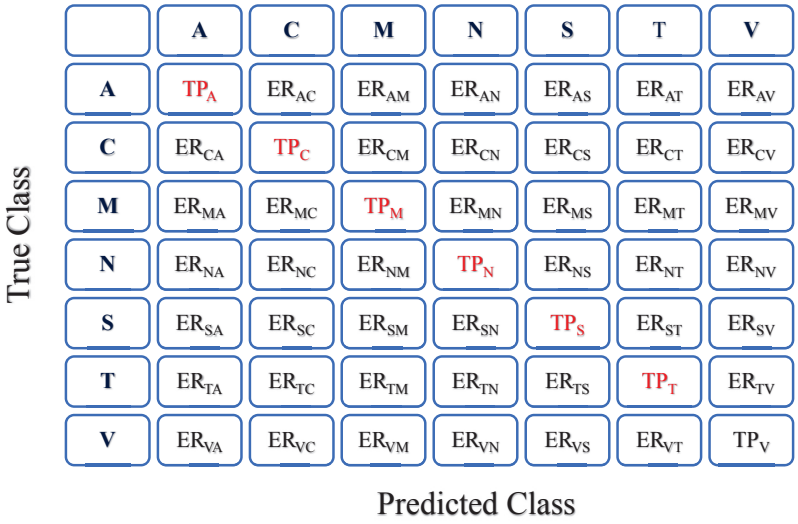


Figure 4.1: Confusion matrix for multi-class audio performance evaluation.

4.4 Performance Comparison

In this section, we discussed the performance comparison of each method used in this thesis work. First, we discussed the performance of three classifier for audio event recognition using MFCC as feature. Then, we discussed the performance of 2D-CNN classifier using spectrogram image as feature for audio event recognition. Finally, we compare the all four AER schemes used in thesis work for audio event recognition.

4.4.1 AER with MFCC Feature

In the first stage of our thesis work, we used MFCC feature to represent the audio and three classifiers based on machine learning and deep learning such as MC-SVM, FFNN and 1D-CNN to detect audio events. While training these models we observed that FFNN took less time for training comparison to MC-SVM and 1D-CNN model. It is

also observed that the trained model size of 1D-CNN is very less contrast to MC-SVM and FFNN models because the number of learning parameters in 1D-CNN is less than other models. The method is evaluated on 70 min of 7 class audio dataset against different frame length. For each frame length, the confusion matrix of three AER schemes are summarized in Table 4.3 and their overall performances are summarized in Table 4.4. The results shows that the FFNN and 1D-CNN based AER schemes had the F1-score values of 95.72% and 96.34% for audio frame size of 250 ms whereas MC-SVM based AER scheme had an overall accuracy of 85.84%. The 1D-CNN based AER scheme had a class-wise accuracy is greater than 84% whereas the FFNN based scheme had a class-wise accuracy is greater than 80%. It is also observed that the recognition rate of aircraft, nature, and train are more than 90%. The computational analysis results show that the prediction time of 1D-CNN based scheme is faster than the FFNN based AER scheme.

4.4.2 AER with Spectrogram Image Feature

In second stage of our thesis work, we used spectrogram image feature to represent audio signal and image classifier 2D-CNN to detect the events. The method is evaluated on 70 min of 7 class audio dataset against different frame length. For each frame length, the confusion matrix of 2D-CNN based AER scheme are summarized in Table 4.5 and their overall performances are summarized in Table 4.6. The result shows that 2D-CNN based AER scheme had the F1-score value of 87.04% and overall accuracy of 88.48% for audio frame size of 100 ms. In this scheme, the class-wise accuracy for 100 ms frame size is greater than 250 ms and 500 ms frame size. It is also observed that the recognition rate of aircraft, nature, construction, train and vehicle are more than 90%.

4.4.3 Comparison of MFCC and Spectrogram Based AER Schemes

The MFCC based AER schemes and spectrogram based AER scheme are evaluated on same dataset of 70 min. It is observed that MFCC with 1D-CNN model with 96.18% overall accuracy outperform the spectrogram based 2D-CNN model with 88.48% overall accuracy. The recognition of aircraft, vehicle, train and nature are the state-of-art of our thesis work.

All four AER schemes were evaluated using a wide variety of audio signals in real-time by using our audio event monitoring graphical user interface (GUI) as shown in Fig.4.2 .

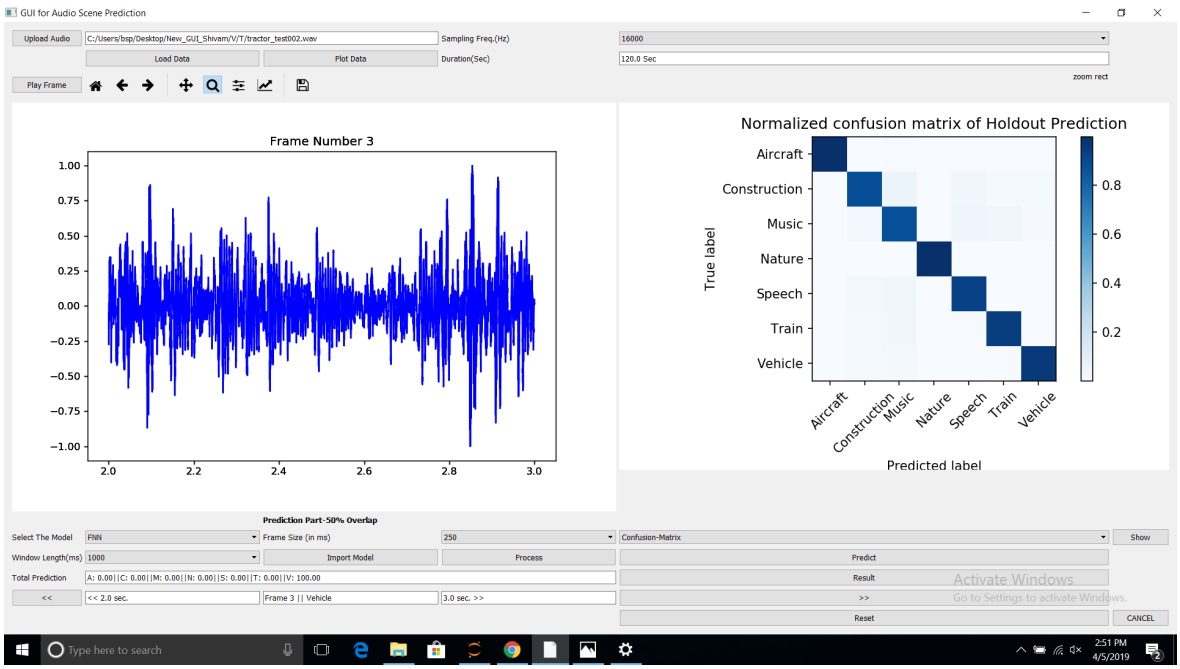
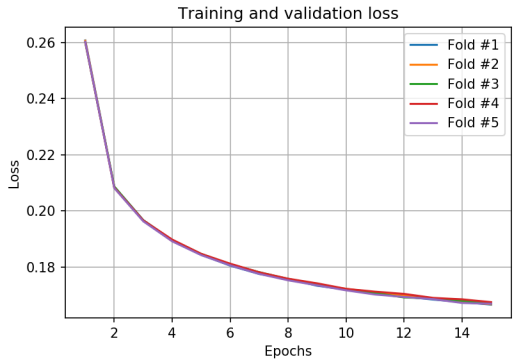
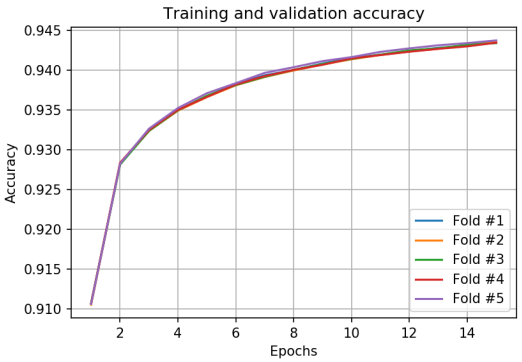


Figure 4.2: Audio event monitoring GUI.

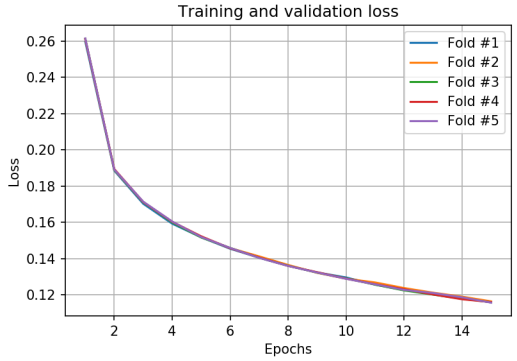


(a) 5-fold CV loss vs epochs

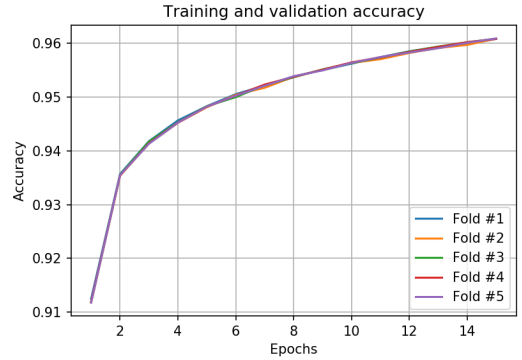


(b) 5-fold CV accuracy vs epochs

Figure 4.3: 5 Fold cross validation curve for 100 ms frame length in FFNN.

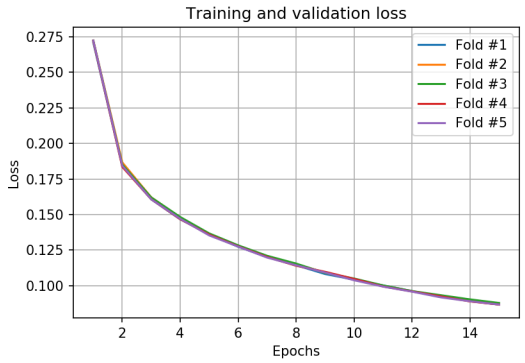


(a) 5-fold CV loss vs epochs

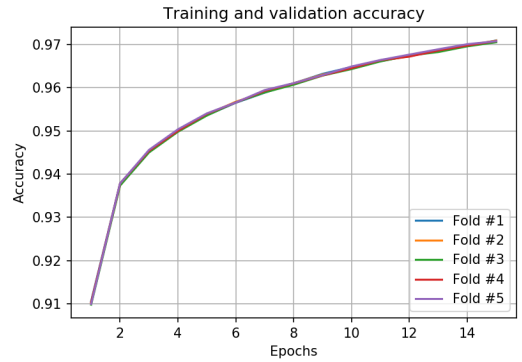


(b) 5-fold CV accuracy vs epochs

Figure 4.4: 5 Fold cross validation curve for 250 ms frame length in FFNN.

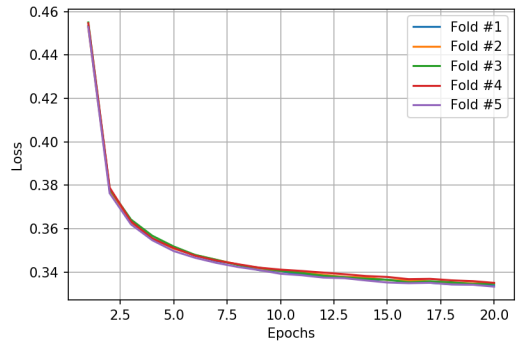


(a) 5-fold CV loss vs epochs

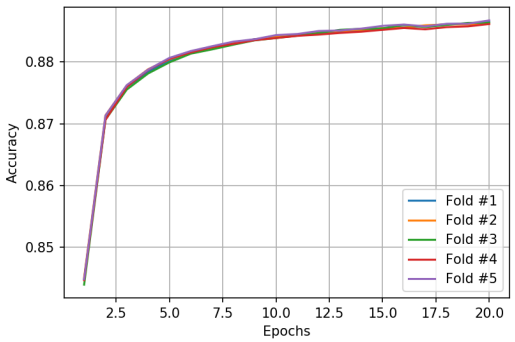


(b) 5-fold CV accuracy vs epochs

Figure 4.5: 5 Fold cross validation curve for 500 ms frame length in FFNN.

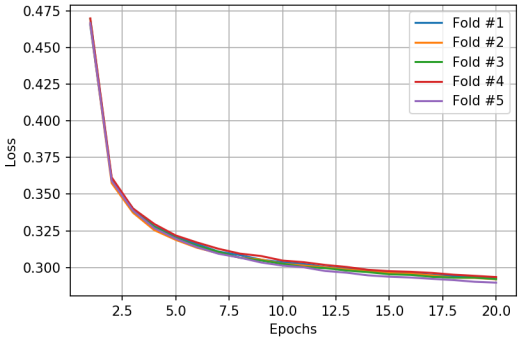


(a) 5-fold CV loss vs epochs

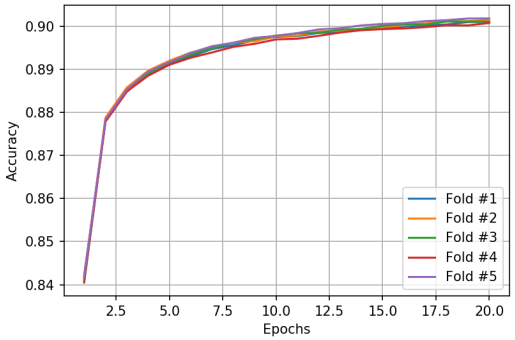


(b) 5-fold CV accuracy vs epochs

Figure 4.6: 5 Fold cross validation curve for 100 ms frame length in 1D-CNN.

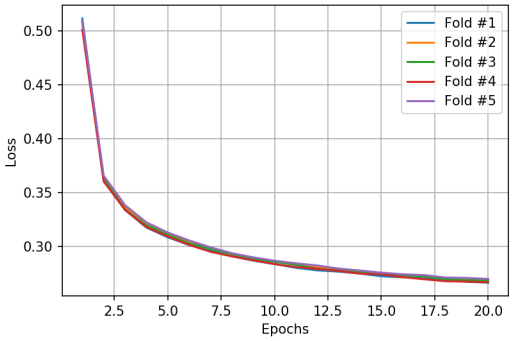


(a) 5-fold CV loss vs epochs

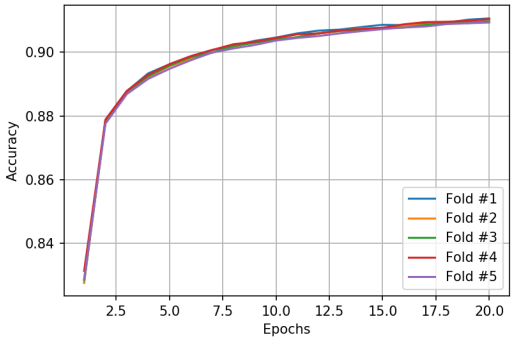


(b) 5-fold CV accuracy vs epochs

Figure 4.7: 5 Fold cross validation curve for 250 ms frame length in 1D-CNN.

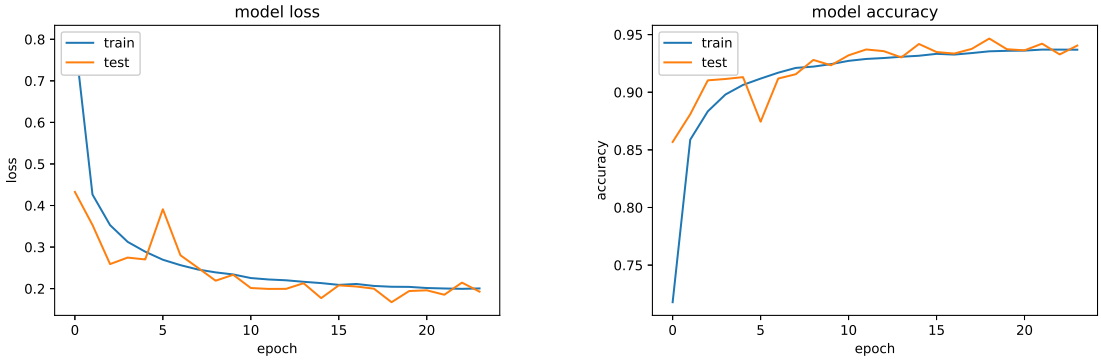


(a) 5-fold CV loss vs epochs



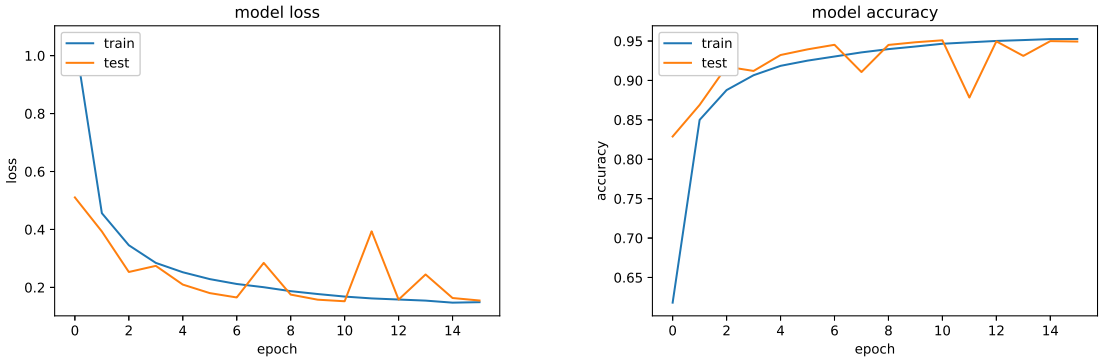
(b) 5-fold CV accuracy vs epochs

Figure 4.8: 5 Fold cross validation curve for 500 ms frame length in 1D-CNN.



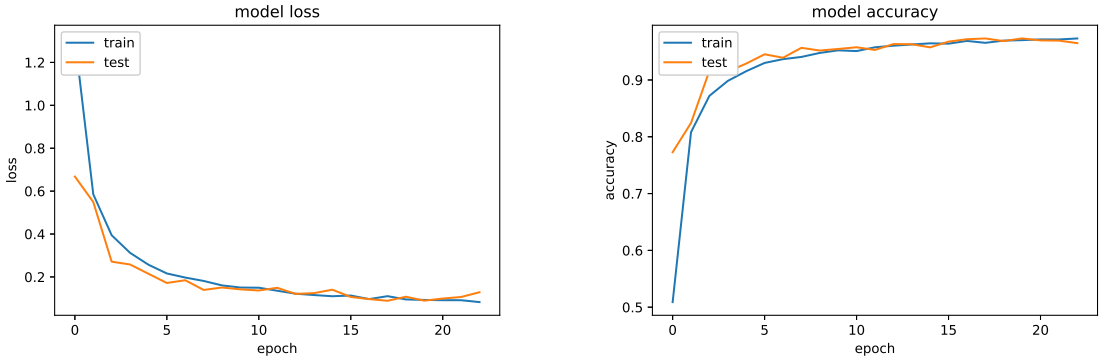
(a) Model loss vs epochs (b) Model accuracy vs epochs

Figure 4.9: 2D-CNN model learning curve for 100 ms frame length.



(a) Model loss vs epochs (b) Model accuracy vs epochs

Figure 4.10: 2D-CNN model learning curve for 250 ms frame length .



(a) Model loss vs epochs (b) Model accuracy vs epochs

Figure 4.11: 2D-CNN model learning curve for 500 ms frame length.

Table 4.2: The comparison of 5-fold accuracy of the three AER schemes using MFCCs feature against different frame length.

Frame Length	**Fold**	**MC-SVM**	**FFNN**	**1D-CNN**
		Accuracy (%)		
100 ms	1	91.74	93.99	90.39
	2	91.54	93.86	90.33
	3	91.62	93.93	90.36
	4	91.3	93.89	90.27
	5	92.2	93.98	90.44
	avg/total	**91.7**	**93.9**	**90.4**
	std.dev.	**0.33**	**0.1**	**0.06**
250 ms	1	92.37	95.02	91.81
	2	92.34	95.11	91.7
	3	92.26	95.15	91.85
	4	92.3	95.15	91.79
	5	92.2	95.17	91.86
	avg/total	**92.3**	**95.1**	**91.8**
	std.dev.	**0.07**	**0.06**	**0.06**
500 ms	1	92.7	95.74	92.51
	2	92.63	95.7	92.33
	3	92.47	95.77	92.65
	4	92.61	95.7	92.64
	5	92.75	95.71	92.68
	avg/total	**92.6**	**95.7**	**92.6**
	std.dev.	**0.11**	**0.03**	**0.15**

Table 4.3: Confusion matrix for three Audio Event Recognition (AER) schemes for frame lengths (FL) of 100 ms, 250 ms, and 500 ms using MFCC as feature

	MC-SVM Based AER							**FFNN Based AER**							**1D-CNN Based AER**						
Audio frame length (FL) = 100 ms																					
	A	**C**	**M**	**N**	**S**	**T**	**V**	**A**	**C**	**M**	**N**	**S**	**T**	**V**	**A**	**C**	**M**	**N**	**S**	**T**	**V**
A	**11838**	12	70	0	20	29	22	**11927**	3	16	0	22	11	12	**11872**	0	36	0	64	16	3
C	72	**3804**	3731	2	169	632	3581	5	**11912**	29	0	36	5	4	3	**11537**	300	0	105	26	20
M	198	497	**10123**	107	237	688	141	51	389	**10598**	18	305	533	97	60	286	**10763**	64	328	416	74
N	0	2	4	**11984**	0	1	0	0	0	6	**11983**	0	2	0	0	3	17	**11969**	0	2	0
S	3	1458	1196	2	**9292**	27	13	1	854	1373	3	**9676**	55	29	5	680	640	2	**10628**	12	24
T	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0	0	0	3	0	0	**11988**	0
V	0	0	0	0	1	0	**11990**	0	0	0	0	0	1	**11990**	0	0	0	0	1	1	**11989**
Audio frame length (FL) = 250 ms																					
A	**11792**	6	45	0	100	19	29	**11954**	5	10	0	14	7	1	**11931**	2	22	0	16	20	0
C	61	**5371**	2434	8	610	228	3279	3	**11940**	29	0	15	2	2	2	**11604**	240	0	56	77	12
M	98	524	**10485**	109	279	416	80	21	413	**10939**	22	219	310	67	23	261	**11223**	18	210	252	4
N	0	0	1	**11990**	0	0	0	0	0	0	**11991**	0	0	0	0	0	4	**11986**	0	1	0
S	10	1638	1090	3	**9196**	44	10	1	1054	1252	0	**9605**	47	32	7	886	883	3	**10179**	11	22
T	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0
V	0	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0	0	1	0	0	0	**11990**
Audio frame length (FL) = 500 ms																					
A	**11779**	0	31	0	141	14	26	**11978**	1	2	0	4	6	0	**11969**	0	1	0	19	2	0
C	125	**5340**	1301	2	1270	195	3758	0	**11899**	26	0	19	44	3	0	**11803**	135	0	46	7	0
M	80	934	**10100**	118	245	367	147	12	536	**10935**	17	171	265	55	3	630	**10590**	33	485	243	7
N	0	0	5	**11986**	0	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0	0	0
S	5	1367	1244	3	**9275**	67	30	6	1399	1503	0	**8965**	58	60	6	1486	1087	1	**9358**	20	33
T	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0
V	0	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**

Table 4.4: Performance of three Audio Event Recognition schemes using MFCC as feature

		MC-SVM			**FFNN**			**1D-CNN**		
Frame Length	**Class**	**Pr**	**Se**	**F1**	**Pr**	**Se**	**F1**	**Pr**	**Se**	**F1**
100 ms	Aircraft	97.74	98.72	98.22	99.52	99.46	99.48	99.43	99	99.21
	Construction	65.89	31.72	42.82	90.53	99.34	94.73	92.25	96.21	94.18
	Music	66.93	84.42	74.66	88.15	88.38	88.26	91.52	89.75	90.62
	Nature	99.08	99.94	99.5	99.82	99.93	99.87	99.45	99.81	99.62
	Speech	95.6	77.49	85.59	96.38	80.69	87.83	95.52	88.63	91.94
	Train	89.69	100	94.56	95.18	100	97.53	96.2	99.97	97.98
	Vehicle	76.14	99.99	86.45	98.82	99.99	99.4	99	99.98	99.48
	Avg.	**84.44**	**84.61**	**83.11**	**95.49**	**95.40**	**95.30**	**96.20**	**96.19**	**96.15**
250 ms	Aircraft	98.58	98.34	98.45	99.79	99.69	99.73	99.73	99.49	99.6
	Construction	71.24	44.54	54.81	89.02	99.57	94.1	90.99	96.77	93.79
	Music	74.59	87.44	80.5	89.44	91.22	90.32	90.7	93.59	92.12
	Nature	99	99.99	99.49	99.81	100	99.9	99.82	99.95	99.88
	Speech	90.28	76.69	82.93	97.48	80.1	87.93	97.3	84.88	90.66
	Train	94.43	100	97.13	97.03	100	98.49	97.07	100	98.51
	Vehicle	77.91	100	87.58	99.15	100	99.57	99.68	99.99	99.83
	Avg.	**86.58**	**86.71**	**85.84**	**95.96**	**95.80**	**95.72**	**96.47**	**96.38**	**96.34**
500 ms	Aircraft	98.24	98.23	98.23	99.84	99.89	99.86	99.92	99.81	99.86
	Construction	69.88	44.53	54.39	86	99.23	92.14	84.79	98.43	91.1
	Music	79.64	84.22	81.86	87.71	91.19	89.41	89.64	88.31	88.97
	Nature	98.98	99.99	99.48	99.85	100	99.92	99.71	100	99.85
	Speech	84.85	77.34	80.92	97.88	74.76	84.77	94.44	78.04	85.46
	Train	94.91	100	97.38	96.98	100	98.46	97.78	100	98.87
	Vehicle	75.16	100	85.81	99.02	100	99.5	99.66	100	99.82
	Avg.	**85.95**	**86.33**	**85.44**	**95.33**	**95.01**	**94.87**	**95.13**	**94.94**	**94.85**

Table 4.5: Confusion matrix for 2D-CNN based Audio Event Recognition (AER) schemes for frame lengths (FL) of 100 ms, 250 ms, and 500 ms using spectrogram as feature

Audio frame length (FL) = 100 ms							
	Aircraft	**Construction**	**Music**	**Nature**	**Speech**	**Train**	**Vehicle**
Aircraft	**11722**	8	60	84	5	108	4
Construction	9	**11920**	26	9	6	6	15
Music	5	565	**9997**	355	206	695	168
Nature	1	2	0	**11982**	0	1	5
Speech	87	2419	3197	65	**4680**	263	1280
Train	0	6	2	0	0	**11981**	2
Vehicle	0	0	0	0	0	0	**11991**
Audio frame length (FL) = 250 ms							
Aircraft	**11061**	15	47	500	5	333	30
Construction	3	**11714**	240	0	6	12	16
Music	0	225	**11270**	85	87	308	16
Nature	0	2	2	**11962**	0	1	24
Speech	7	3179	7008	14	**630**	96	1057
Train	0	0	4	7	0	**11979**	1
Vehicle	0	0	0	0	0	0	**11991**
Audio frame length (FL) = 500 ms							
Aircraft	**11674**	14	6	95	3	185	14
Construction	1	**11918**	44	0	5	14	9
Music	0	198	**10793**	21	124	818	37
Nature	0	0	0	**11990**	0	0	1
Speech	50	2511	4621	8	**2945**	276	1580
Train	0	2	0	0	0	**11989**	0
Vehicle	0	0	0	0	0	0	**11991**

Table 4.6: Performance of 2D-CNN scheme using spectrogram feature

Frame Length	**Class**	**Pr**	**Se**	**F1**
100 ms	Aircraft	99.13	97.75	98.43
	Construction	79.89	99.4	88.58
	Music	75.26	83.37	79.1
	Nature	95.89	99.92	97.86
	Speech	95.56	39.02	55.41
	Train	91.78	99.91	95.67
	Vehicle	89.05	100	94.2
	avg / total	**89.51**	**88.48**	**87.04**
250 ms	Aircraft	99.9	92.24	95.91
	Construction	77.39	97.68	86.35
	Music	60.68	93.98	73.74
	Nature	95.17	99.75	97.4
	Speech	86.53	5.25	9.89
	Train	94.01	99.89	96.86
	Vehicle	91.29	100	95.44
	avg / total	**86.42**	**84.11**	**79.37**
500 ms	Aircraft	99.56	97.35	98.44
	Construction	81.39	99.39	89.49
	Music	69.79	90	78.61
	Nature	98.97	99.99	99.47
	Speech	95.71	24.56	39.08
	Train	90.26	99.98	94.87
	Vehicle	87.96	100	93.59
	avg / total	**89.09**	**87.32**	**84.79**

Chapter 5

Conclusions and Future Directions

In this thesis, we presented four audio sound recognition (AER) schemes using the MFCC and spectrogram as audio features with machine learning and deep learning based classifiers such as multi-class SVM, FFNN, 1D-CNN, and 2D-CNN for recognizing 7 sound classes. We created large scale audio database including aircraft, construction, music, nature (wind and rain), speech, vehicle, and train sounds by considering different audio recording devices such as commercial hand-held audio recorders, smartphones, laptop PC and different recording environments. In the first stage of our thesis work, we used MFCC as feature and MC-SVM, FFNN, 1D-CNN as classifiers to classify the seven class audio events. The results showed that 1D-CNN outperforms other schemes for audio frame length 250 ms. In the final stage of our thesis work, we used spectrogram image as feature and 2D-CNN as classifier to classify the seven class audio events. The results showed that recognition rate for 100 ms frame length based 2D-CNN classifier is more than 250 ms and 500 ms. We also observed that MFCC feature with 1D-CNN classifier gives more recognition accuracy than spectrogram feature based 2D-CNN classifier. All four AER schemes were evaluated using a wide variety of audio signals in real-time by using our audio event monitoring graphical user interface (GUI). The results demonstrated the significance of selection of frame length for achieving better recognition performance. The recognition of aircraft, train and vehicle sounds with more than 90% accuracy are the state-of-art of our thesis work.

For the future work, collection of more audio database for testing of our methods. We can also do the testing of our classifier models with available standard audio database. There is also possibility improving the performance of models by applying post processing method and reduce the computational complexity in prediction time by studying the hyper-parameter and algorithm of classifier. The frame wise multi-label classification is the future research in this area.

Publications

Conference Publication

1. S. Soni, S. Dey and M. S. Manikandan ,"Automatic audio event recognition schemes for context aware audio computing devices," in *Proc. 7th Int. Conf. on Digital Information Processing and Communications (ICDIPC)*, Turkey, 2019. (accepted)

References

[1] R. S. Goldhor, "Recognition of environmental sounds," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,* vol. 1, pp. 149-152, 1993.

[2] Jia-Ching Wang, Jhing-Fa Wang, Kuok Wai He and Cheng-Shu Hsu, "Environmental Sound Classification using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor," in *Proc. IEEE Int. Joint Conf. on Neural Network Proceedings,* pp. 1731-1735, 2006.

[3] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters,* Elsevier, vol. 24, no. 15, pp. 2895–2907, 2003.

[4] F. Beritelli and R. Grasso, "A pattern recognition system for environmental sound classification based on MFCCs and neural networks," in *Proc. 2nd Int. Conf. on Signal Processing and Communication Systems,* pp. 1-4, 2008.

[5] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen and T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 6, pp. 1291-1303, 2017.

[6] S. Souli and Z. Lachiri, "Audio sounds classification using scattering features and support vectors machines for medical surveillance," *Applied Acoustics,* Elsevier, vol. 130, pp. 270 - 282, 2018.

[7] K. Hwang and S. Lee, "Environmental audio scene and activity recognition through mobile-based crowd sourcing," *IEEE Transactions on Consumer Electronics,* vol. 58, no. 2, pp. 700-705, 2012.

[8] X. Valero and F. Alias, "Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification," *IEEE Transactions on Multimedia,* vol. 14, no. 6, pp. 1684-1689, 2012.

[9] C. Lee, C. Han and C. Chuang, "Automatic Classification of Bird Species From Their Sounds Using Two-Dimensional Cepstral Coefficients," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 16, no. 8, pp. 1541-1550, 2008.

[10] B. Ghoraani and S. Krishnan, "Time-Frequency Matrix Feature Extraction and Classification of Environmental Audio Signals," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 7, pp. 2197-2209, 2011.

[11] Chien-Chang Lin, Shi-Huang Chen, Trieu-Kien Truong and Yukon Chang, "Audio classification and categorization based on wavelets and support vector Machine,"

IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 644-651, 2005.

[12] N. Almaadeed, M. Asim, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "Automatic Detection and Classification of Audio Events for Road Surveillance Applications," *Sensors (Basel)*, vol. 18, no. 6, 2018.

[13] S. Chu, S. Narayanan and C.C. J. Kuo, "Environmental sound recognition using MP-based features," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* pp. 1-4, 2008.

[14] S. Chu, S. Narayanan and C.C. J. Kuo, "Environmental Sound Recognition With Time- Frequency Audio Features," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17, no. 6, pp. 1142-1158, 2009.

[15] Uzkent B., Barkana B.D. and Cevikalp H., "Non-speech environmental sound classification using SVMs with a new set of features", *Int. Journal of Innovative Computing, Information and Control,* vol.8, no. 5, pp.3511-3524, 2012.

[16] J. Wang, C. Lin, B. Chen and M. Tsai, "Gabor-Based Nonuniform Scale-Frequency Map for Environmental Sound Classification in Home Automation," *IEEE Transactions on Automation Science and Engineering,* vol. 11, no. 2, pp. 607-613, 2014.

[17] J. Wang, Y. Lee, C. Lin, E. Siahaan and C. Yang, "Robust Environmental Sound Recognition With Fast Noise Suppression for Home Automation," *IEEE Transactions on Automation Science and Engineering,* vol. 12, no. 4, pp. 1235-1242, 2015.

[18] A. Rakotomamonjy, "Supervised Representation Learning for Audio Scene Classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 6, pp. 1253-1265, 2017.

[19] H. D. Tran and H. Li, "Sound Event Recognition With Probabilistic Distance SVMs," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 6, pp. 1556-1568, 2011.

[20] I. McLoughlin, H. Zhang, Z. Xie, Y. Song and W. Xiao, "Robust Sound Event Classification Using Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 3, pp. 540-552, 2015.

[21] W. Yang and S. Krishnan, "Combining Temporal Features by Local Binary Pattern for Acoustic Scene Classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 6, pp. 1315-1321, 2017.

[22] Rabaoui, Asma, Manuel Davy, Stephane Rossignol, and Noureddine Ellouze, "Using one-class SVMs and wavelets for audio surveillance," *IEEE Transactions on information forensics and security ,* vol. 3, no. 4, pp.763-775, 2008.

[23] S. Sigtia, A. M. Stark, S. Krstulović and M. D. Plumbley, "Automatic Environmental Sound Recognition: Performance Versus Computational Cost," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 11, pp. 2096-2107, 2016.

[24] O. Gencoglu, T. Virtanen and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *Proc. 22nd European Signal Processing Conference (EUSIPCO),* pp. 506-510, 2014.

[25] S. Souli and Z. Lachiri, "Environmental sound classification using log-Gabor filter," in *Proc. IEEE 11th Int. Conf. on Signal Processing,* pp. 144-147, 2012.

[26] H. Zhang, I. McLoughlin and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proc.IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* pp. 559-563, 2015.

[27] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," emphProcedia Computer Science, Elsevier, vol. 112, pp. 2048–2056, 2017.

[28] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events," *IEEE Transactions on Multimedia,* vol. 17, no. 10, pp. 1733-1746, 2015.

[29] A. Mesaros et al., "DCASE 2017 Challenge setup: Tasks, datasets and baseline system," in *Proc. DCASE 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events,* 2017.

[30] H. Phan, P. Koch, F. Katzberg, M. Maass, R. Mazur, and A. Mertins, "Audio Scene Classification with Deep Recurrent Neural Networks," *arXiv:1703.04770* [cs], 2017.

[31] H. Phan et al., "What Makes Audio Event Detection Harder than Classification?," in *Proc. 25th European Signal Processing Conf. (EUSIPCO),* pp. 2739–2743, 2017.

[32] S. Abidin, R. Togneri and F. Sohel, "Spectrotemporal Analysis Using Local Binary Pattern Variants for Acoustic Scene Classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 26, no. 11, pp. 2112-2121, 2018.

[33] Y. Kim, M. Kim, J. Goo and H. Kim, "Learning Self-Informed Feature Contribution for Deep Learning-Based Acoustic Modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 26, no. 11, pp. 2204-2214, 2018.

[34] C. Wang, J. Wang, A. Santoso, C. Chiang and C. Wu, "Sound Event Recognition Using Auditory-Receptive-Field Binary Pattern and Hierarchical-Diving Deep Belief Network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 26, no. 8, pp. 1336-1351, 2018.

[35] A. Mesaros et al., "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 26, no. 2, pp. 379-393, 2018.

[36] G. Richard, T. Virtanen, J. P. Bello, N. Ono and H. Glotin, "Introduction to the Special Section on Sound Scene and Event Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 6, pp. 1169-1171, 2017.

[37] J. Schroder, N. Moritz, J. Anemuller, S. Goetze and B. Kollmeier, "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, 2017.

[38] N. R. Koluguri, G. N. Meenakshi and P. K. Ghosh, "Spectrogram Enhancement Using Multiple Window Savitzky-Golay (MWSG) Filter for Robust Bird Sound Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,*vol. 25, no. 6, pp. 1183-1192, 2017.

[39] V. Bisot, R. Serizel, S. Essid and G. Richard, "Feature Learning With Matrix Factorization Applied to Acoustic Scene Classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 6, pp. 1216-1229, 2017.

[40] K. Imoto and N. Ono, "Spatial Cepstrum as a Spatial Feature Using a Distributed Microphone Array for Acoustic Scene Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 6, pp. 1335-1343, 2017.

[41] I. Trowitzsch, J. Mohr, Y. Kashef and K. Obermayer, "Robust Detection of Environmental Sounds in Binaural Auditory Scenes," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 6, pp. 1344-1356, 2017.

[42] H. Phan, L. Hertel, M. Maass, R. Mazur and A. Mertins, "Learning Representations for Nonspeech Audio Events Through Their Similarities to Speech Patterns," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 4, pp. 807-822, 2016.

[43] Sharan, Roneel V., and Tom J. Moir, "Acoustic event recognition using cochleagram image and convolutional neural networks," *Applied Acoustics* vol. 148, pp. 62-66., 2019.

[44] Z. Ren et al., "Deep Scalogram Representations for Acoustic Scene Classification," *IEEE/CAA Journal of Automatica Sinica,* vol. 5, no. 3, pp. 662-669, 2018.

[45] T. Kobayashi and J. Ye, "Acoustic feature extraction by statistics based local binary pattern for environmental sound classification," in*Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* pp. 3052-3056, 2014.

[46] J. Ye, T. Kobayashi, and M. Murakawa, "Urban sound event classification based on local and global features aggregation," *Applied Acoustics,* Elsevier, vol. 117, pp. 246–256, 2017.

[47] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters,* vol. 24, no. 3, pp. 279-283, 2017.

[48] I. Ozer, Z. Ozer, and O. Findik, "Noise robust sound event classification with convolutional neural network," *Neurocomputing,* Elsevier, vol. 272, pp. 505–512, 2018.

[49] I. Ozer, Z. Ozer, and O. Findik, "Lanczos kernel based spectrogram image features for sound classification," *Procedia Computer Science,* Elsevier, vol. 111, pp. 137–144, 2017.

[50] Yong Xu, Qiang Huang, Wenwu Wang, Mark D. Plumbley MD, "Hierarchical learning for DNN-based acoustic scene classification," *arXiv preprint arXiv:1607.03682,* 2016.

[51] J. Li, W. Dai, F. Metze, S. Qu and S. Das, "A comparison of Deep Learning methods for environmental sound detection," in*Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* pp. 126-130, 2017.

[52] A. Mesaros, T. Heittola and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in*Proc. 24th European Signal Processing Conference (EUSIPCO),* pp. 1128-1132, 2016.

[53] D. Barchiesi, D. Giannoulis, D. Stowell and M. D. Plumbley, "Acoustic Scene Classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16-34, 2015.

[54] P. Khunarsal, C. Lursinsap, and T. Raicharoen, "Very short time environmental sound classification based on spectrogram pattern matching," *Information Sciences,* Elsevier, vol. 243, pp. 57–74, 2013.

[55] J. Dennis, H. D. Tran and E. S. Chng, "Image Feature Representation of the Subband Power Distribution for Robust Sound Event Classification," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 2, pp. 367-377, 2013.

[56] J. Wang, H. Lee, J. Wang and C. Lin, "Robust Environmental Sound Recognition for Home Automation," *IEEE Transactions on Automation Science and Engineering,* vol. 5, no. 1, pp. 25-31, 2008.

[57] G. Wichern, J. Xue, H. Thornburg, B. Mechtley and A. Spanias, "Segmentation, Indexing, and Retrieval for Environmental and Natural Sounds," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, no. 3, pp. 688-707, 2010.

[58] James lyons, "`python_speech_features`:common speech features for ASR, 2013-", *https://github.com/jameslyons/`python_speech_features`,* [Online; accessed 2019-01-15].

[59] Pedregosa, F., Varoquaux, G., Gramfort, A. & Michel, V., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[60] L. Buitinck, G. Louppe, M. B. a. Pedregosa, A. Mueller, O. G. a. Niculae, P. Prettenhofer, A. G. J. Grobler, R. Layton, J. V. a. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," *ECML PKDD Workshop: Languages for Data Mining and Machine Learning,* pp. 108-122, 2013.

[61] Brian McFee, Matt McVicar, Stefan Balke, Vincent Lostanlen, Carl Thome, Colin Raffel, CJ Carr., "librosa/librosa: 0.6.3 (Version 0.6.3). Zenodo.," *http://doi.org/10.5281/zenodo.2564164*, 2019.

[62] S. Waldekar and G. Saha, "Classification of audio scenes with novel features in a fused system framework," *Digital Signal Processing,* Elsevier, vol. 75, pp. 71–82, 2018.

[63] Ye, Jiaxing, Takumi Kobayashi, Nobuyuki Toyama, Hiroshi Tsuda, and Masahiro Murakawa, "Acoustic Scene Classification Using Efficient Summary Statistics and Multiple Spectro-Temporal Descriptor Fusion," *Applied Sciences* vol 8, no. 8, pp. 1363, 2018.

[64] R. V. Sharan and T. J. Moir, "Robust acoustic event classification using deep neural networks," *Information Sciences,* Elsevier, vol. 396, pp. 24–32, Aug. 2017.

[65] F. Medhat, D. Chesmore and J. Robinson, "Recognition of Acoustic Events Using Masked Conditional Neural Networks," in*Proc. 16th IEEE Int. Conf. on Machine Learning and Applications (ICMLA),* pp. 199-206, 2017.

[66] G. Takahashi, T. Yamada, N. Ono and S. Makino, "Performance evaluation of acoustic scene classification using DNN-GMM and frame-concatenated acoustic features," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),* pp. 1739-1743, 2017.

[67] K. Z. Thwe and N. War, "Environmental sound classification based on time-frequency representation," in *Proc. 18th IEEE/ACIS Int. Conf. on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD),* pp. 251-255, 2017.

[68] M. G. Lopez-Pacheco, L. P. Sánchez-Fernández, H. Molina-Lozano, and L. A. Sanchez-Pérez, "Predominant environmental noise classification over sound mixing based on source-specific dictionary," *Applied Acoustics,* Elsevier, vol. 112, pp. 171–180, 2016.

[69] L. Hertel, H. Phan and A. Mertins, "Comparing time and frequency domain for audio event recognition using deep learning," in *Proc.Int. Joint Conf. on Neural Networks (IJCNN),* pp. 3407-3411, 2016.

[70] Crocco, Marco, Marco Cristani, Andrea Trucco, and Vittorio Murino, "Audio surveillance: a systematic review" in *ACM Computing Surveys (CSUR)* vol. 48, no. 4, pp.52, 2016.

[71] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in*Proc. IEEE 25th Int. Workshop on Machine Learning for Signal Processing (MLSP),* pp. 1-6, 2015.

[72] W. Hsieh, C. Ho, V. Duong, Y. Lee and J. Wang, "2D semi-NMF of scale-frequency map for environmental sound classification," in *Proc. Signal and Information Processing Association Annual Summit and Conf.(APSIPA),* pp. 1-4, 2014.

[73] M. Karbasi, S. M. Ahadi and M. Bahmanian, "Environmental sound classification using spectral dynamic features," in *Proc. 8th Int. Conf. on Information, Communications & Signal Processing,* pp. 1-5, 2011

[74] J. R. Delgado-Contreras, J. P. Garcıa-Vazquez, R. F. Brena, C. E. Galvan-Tejada, and J. I. Galvan-Tejada, "Feature Selection for Place Classification through Environmental Sounds," *Procedia Computer Science,* Elsevier, vol. 37, pp. 40–47, 2014.

[75] C. Bauge, M. Lagrange, J. Anden and S. Mallat, "Representing environmental sounds using the separable scattering transform," in*IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* pp. 8667-8671, 2013.

[76] S. Sivasankaran and K. M. M. Prabhu, "Robust features for environmental sound classification," in *IEEE Int. Conf. on Electronics, Computing and Communication Technologies,* pp. 1-6, 2013.

[77] G. Shen, Q. Nguyen, and J. Choi, "An Environmental Sound Source Classification System Based on Mel-Frequency Cepstral Coefficients and Gaussian Mixture Models," *IFAC Proceedings Volumes,* Elsevier, vol. 45, no. 6, pp. 1802–1807, 2012.

[78] K. Wang, L. W. Yang, and B. Yang, "Audio Event Detection and classification using extended R-FCN Approach," 2017.

[79] Heittola, Toni, Annamaria Mesaros, Tuomas Virtanen and Antti J. Eronen, "Sound Event Detection and Context Recognition," 2011.

[80] Singh, Arshdeep, et al. "A Layer-wise Score Level Ensemble Framework for Acoustic Scene Classification," 2017.

[81] Maxime, J., Alameda-Pineda, X., Girin, L. and Horaud, R., 2014, "Sound representation and classification benchmark for domestic robots," in *Proc IEEE Int. Conf. on Robotics and Automation (ICRA),* pp. 6285-6292, 2014.

[82] Chollet, F., "keras," GitHub. https://github.com/fchollet/keras, 2015.

[83] H. G. Okuno and K. Nakadai, "Computational Auditory Scene Analysis and its Application to Robot Audition," *Hands-Free Speech Communication and Microphone Arrays,* pp. 124-127, 2008.

[84] Chollet, Francois, "Deep learning with python," *Manning Publications Co.,* 2017.

[85] Lerch, Alexander, "An introduction to audio content analysis: Applications in signal processing and music informatics" *Wiley-IEEE Press.*, 2012.

[86] V. N. Vapnik, "The Nature of Statistical Learning Theory," Springer-Verlag, London, UK 1995.

[87] V. N. Vapnik, "Statistical Learning Theory," John Wiley & Sons, New York, NY, 1998.

[88] C. Liu, Y. Yang and C. Tang, "An Improved Method for Multi-class Support Vector Machines," in *Proc. Int. Conf. on Measuring Technology and Mechatronics Automation,* pp. 504-508, 2010.

[89] Chih-Wei Hsu and Chih-Jen Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks,* vol. 13, no. 2, pp. 415-425, 2002.

[90] Abe, Shigeo, "Analysis of multiclass support vector machines," in *Proc. Int. Conf. on Computational Intelligence for Modelling Control and Automation (CIMCA),* vol. 21, no. 3, pp. 385-396, 2003.

[91] Kijsirikul B. and Ussivakul N., "Multiclass support vector machines using adaptive directed acyclic graph," in *Proc. of Int. Joint Conf. on Neural Networks (IJCNN),* pp. 980-985, 2002.

[92] Chang, C.C. and Lin, C.J., "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.,* 2011.