Table of Contents

List of Figures v

List of Tables vii

Certificate ix

Acknowledgements xi

Abstract xiii

1 Introduction 1

1.1 Acoustics 1

1.1.1 Applications of Acoustics 1

1.2 Acoustic Classification Systems 3

1.2.1 Acoustic Scene Classification 3

1.2.2 Acoustic Event Classification 3

1.3 Traffic event detection using acoustics 4

1.4 Motivation 4

1.5 Objectives 6

1.6 Organization Of Thesis 6

2 Literature Survey 9

2.1 Static Sensors Based Techniques 10

2.1.1 Intrusive Sensing Techniques 10

2.1.2 Non-Intrusive Sensing Techniques 13

2.2 On-Vehicle Sensor Based Techniques 18

2.2.1 Global Positioning System (GPS) 18

2.2.2 Accelerometer Sensors 19

2.3 Basics of Acoustic Processing 21

2.3.1 Basic Principles of Sound 21

2.3.2 Feature Extraction . . . 27
2.3.3 Classification Techniques . . . 33
2.4 Convolutional Neural Networks . . . 37
2.4.1 Architecture of Convolutional Neural Network . . . 37
2.4.2 CNN in Acoustics . . . 40
2.5 Research Gaps . . . 40

3 Methodology **43**
3.1 System Overview . . . 43
3.1.1 Preprocessing . . . 44
3.1.2 Feature Extraction . . . 44
3.1.3 Classification . . . 45
3.2 Approaches Compared . . . 46
3.2.1 Human-defined features based machine learning approach (MFCCs and SVM) . . . 46
3.2.2 Human-defined features based deep learning approach (MFCCs and CNN) . . . 47
3.2.3 Proposed off the shelf CNN features based approach (CNN and SVM) . . . 48
3.3 Determination of efficacy of CNN for feature extraction . . . 49
3.3.1 Model 1: 2-layered CNN Model and classification using SVM . . . 49
3.3.2 Model 2: 4-layered CNN Model and SVM . . . 51
3.4 Feature Fusion . . . 52
3.5 Summary . . . 53

4 Results And Discussions **55**
4.1 Introduction . . . 55
4.2 Capturing Sounds of Vehicles . . . 56
4.2.1 Types Of Vehicles . . . 56
4.2.2 Dataset . . . 56
4.2.3 Dataset Characterstics . . . 57
4.2.4 Experimental Setup . . . 57
4.3 Performance Metrics . . . 57
4.3.1 True Positive(TP) . . . 58
4.3.2 False Positive(FP) . . . 58
4.3.3 True Negative(TN) . . . 58
4.3.4 False Negative(FN) . . . 58

4.3.5 Accuracy . . . 58
4.3.6 Precision . . . 59
4.3.7 Recall . . . 59
4.3.8 F1 score . . . 59
4.4 Comparative Analysis of three approaches . . . 59
4.4.1 Human-defined features based machine learning approach (MFCCs and SVM) . . . 59
4.4.2 Human-defined features based deep learning approach (MFCCs and CNN) . . . 60
4.4.3 Proposed off the shelf CNN features based approach (CNN and SVM) . . . 60
4.4.4 Comparative Analysis of Model 1 and Model 2 of the proposed approach . . . 69
4.5 Performance comparison of three approaches . . . 70
4.6 Feature Fusion . . . 70
4.6.1 Effect of increase in stride length . . . 70
4.6.2 Effect of Increase in number of filters . . . 71
4.6.3 Effect of pool size in max-pooling layer . . . 71
4.6.4 Effect of Size of filter in convolutional layer . . . 72
4.6.5 Analysis of Various kernel functions . . . 72
4.7 Analysis of Feature Fusion . . . 74
4.8 Summary . . . 74

5 Conclusion **75**
5.1 Scope for future work . . . 76

References **79**

List of Figures

1.1 Acoustic Scene Classification 3

2.1 Traffic Sensing Technologies 9
2.2 Intrusive Sensing Techniques 10
2.3 Inductive Loop Detector 11
2.4 Single Element Piezoelectric Sensor 12
2.5 Non-Intrusive Sensing Techniques 13
2.6 Movement Based Sensing Techniques 18
2.7 Rectangular Window 24
2.8 Hanning Window 25
2.9 Hamming Window 25
2.10 Blackman Window 26
2.11 Hyperplane in SVM 35
2.12 Non-Linear Mapping 35
2.13 Convolutional Neural Network 37

3.1 System Overview of Proposed approach 44
3.2 Approaches Compared 46
3.3 Human-defined features based machine learning approach (MFCCs and SVM) 47
3.4 Human-defined features based deep learning approach (MFCCs and CNN) 48
3.5 Proposed off the shelf CNN features based approach (CNN and SVM) . . 49
3.6 Feature Extraction using CNN 50
3.7 4-layered CNN model 51
3.8 Feature Fusion 53

4.1 Classes of Vehicles 56
4.2 Confusion Matrix 58
4.3 Effect of number of filters on the performance of model (Model 1) 61

4.4 Effect of stride length (Model 1) 62
4.5 Analysis of different kernels of SVM (Model 1) 63
4.6 Effect of pool size in max-pooling layer on classification accuracy (Model 1) 64
4.7 Effect of size of input on classification accuracy (Model 1) 64
4.8 Effect of size of filters on classification accuracy (Model 1) 65
4.9 Effect of change in Stride length on performance (Model 2) 66
4.10 Effect of size of filters on classification accuracy (Model 2 setup 1) 67
4.11 Analysis of kernel functions of SVM (Model 2 setup 1) 68
4.12 Effect of number of filters on performance (Model 2 setup 1) 69
4.13 Comparative Analysis of three approaches 70
4.14 Effect of stride on classification accuracy (Feature Fusion) 71
4.15 Effect of Number of filters on performance of model (Feature Fusion) . . 72
4.16 Effect of pool size on performance of model (Feature Fusion) 72
4.17 Effect of size of filter on classification accuracy (Feature Fusion) 73
4.18 Analysis of kernel functions of SVM (Feature Fusion) 73

List of Tables

2.1 Determination of Traffic Parameters Using Acoustics 16
2.2 Sensing Technologies for determining Traffic parameters 19

3.1 Hyperparameters for Model 2 52
3.2 Hyperparameters for Feature Fusion 53

4.1 Number of recordings for each category of vehicle in dataset 56
4.2 Vehicle categories and respective labels used for experiments 57
4.3 Classification Accuracy in approach MFCCs and SVM 60
4.4 Number of features corresponding to number of filters and stride length (Model 1) 61
4.5 Effect of Stride length on classification accuracy (Model 1) 62
4.6 Analysis of different kernels of SVM (Model 1) 63
4.7 Best combination of hyperparameters for Model 1 65
4.8 Number of features corresponding to number of filters and stride length (Model 2) 66
4.9 Effect of stride length on classification accuracy (Model 2) 67
4.10 Analysis of kernel functions of SVM (Model 2 Setup1) 68
4.11 Best combination of hyperparameters for Model 2 69
4.12 Effect of stride length on classification accuracy (Feature Fusion) 71
4.13 Analysis of kernel functions of SVM (Feature Fusion) 73

Certificate

Department of Chemical Engineering

Indian Institute of Technology, Bombay

The report type entitled "Essential LaTeX Templates for Report Writing" submitted by My name (Roll No. ....) may be accepted for being evaluated.

Date: 28 June 2018

Supervisor name

Acknowledgements

Foremost, I bow to the almighty for providing me the strength to carry out my research work with sincerity and dedication. I express my deep sense of gratitude to my dissertation supervisor *Dr. Naveen Aggarwal*, Associate Professor, UIET, Panjab University for his continuous support, guidance, patience and motivation throughout the research work. His constructive criticism and continuous assistance in all my endeavours brought the best out of me. The work could not reach its culmination without critical analysis and valuable suggestions of the co-supervisor *Dr. Akashdeep*, Assistant Professor, UIET, Panjab University. I would like to thank him for his pedagogical guidance and enthusiasm. Furthermore, I want to express my thanks to *Mr. Dinesh Vij*, PhD Scholar, UIET, Panjab University for his help and support.

Special thanks to *Prof. Harish Kumar*, Professor, UIET, Panjab University and *Prof. Savita Gupta*, Director UIET, Panjab University for encouraging good quality of research in the department.

I express my profound gratitude to my family and friends for providing me the support and encouragement throughout my years of study and the process of research.

Anam Bansal
UIET, Panjab University
28 June 2018

Abstract

Recognizing the type of vehicles on the road is very important for traffic management policy makers, insurance companies, public safety organisations etc. Further, classification of vehicles plying on the road can aid in the number of applications like traffic modeling, parking management, toll setting, vehicle identification and surveillance, road maintenance, emissions/pollution estimation etc. Different techniques proposed for vehicle classification can be categorized as infrastructure-based and infrastructureless techniques. Infrastructure-based solutions based on magnetic sensors and inductive loop detectors are costly and entail huge maintenance and installation charges. Video cameras based solutions are affected by occlusion, adverse weather and light conditions. Infrastructureless techniques for vehicle classification mainly rely on smart phone sensors such as using GPS, accelerometers etc. but these techniques are not tested in varied traffic conditions. Then, there are acoustics based techniques for vehicle classification which mainly use human-defined features. In this thesis, we have explored the possibility of using the Convolutional Neural Network(CNN) for extracting features from sounds of vehicles. An off the shelf CNN based feature extraction approach is proposed. These features are used for classification of vehicles using Support Vector Machine (SVM).

The proposed model is trained on a dataset of 4789 recordings collected using commuters' smartphone for five popular public transport vehicles -car, bus, aeroplanes, trains and three-wheeler. The data for these vehicle categories is acquired through commuter's smartphone. Each recording is of size 30 seconds, which is further preprocessed and divided into frames of size 8192 samples. Then CNN model is used for feature extraction and classification is performed through SVM. This approach is compared with Mel Frequency Cepstral Coefficients(MFCCs) features based machine learning and deep learning approach. In proposed approach, two models- two-layered CNN model and four-layered CNN model are used for experiments to determine if the change in number of layers results in increased performance. Number of filters, dimensions of filters, stride length, dimensions of pool size and different kernel functions are analyzed to determine the best combination of hyperparameters. The number of features extracted from CNN vary with

the change in number of filters and stride length. There is marginal difference in the classification accuracy in cases when two-layered CNN model is used with SVM or four-layered CNN model is used with SVM. The best results (99.06%) are obtained with two-layered CNN model with 8 filters, each of size 9*9, stride length of 2*2 in convolutional layer, pool size of 9*9 in max-pooling layer and linear or sigmoid kernel function in SVM. In this case, 32 features are extracted using CNN. The proposed approach is considerably better than MFCCs based machine learning approach (72.65%) and MFCCs based deep learning approach (79.97%).

Further, experiments are performed to determine if the combination of human-defined features and features extracted from CNN could yield good results. 13 MFCCs are combined with 32 features extracted from CNN. Experiments are performed with the different number of filters, stride length, dimensions of filters, pool size in max-pooling layer and different kernel functions. The feature fusion do not yield better results than the approach using off the shelf CNN features. The best classification accuracy of 82.15% is obtained with the 8 filters each of dimension 9*9, stride length of 4*4, pool size of 9*9 in max-pooling layer and Radial Basis function in SVM.

Chapter 1

Introduction

1.1 Acoustics

Acoustics is a branch of physics which is concerned with the study of sound. Sound can be described as mechanical waves in gases, liquids, and solids. In other words, Acoustics is the science of sound concerned with the production, control, transmission, reception, and effects of sound. Acoustics have helped in a number of applications in everyday life like in architectural design of concert halls, context classification, gunshot detection, wildlife monitoring etc. Representation of sound as an electric voltage is an acoustic signal. Humans can hear sound in the range of 20Hz to 20 kHz.

1.1.1 Applications of Acoustics

Acoustics are employed in number of fields. Some of the applications of acoustics are described below:

1. **Speech:** Speech Processing refers to the study of speech and its methods for processing. Speech Processing finds number of applications in speech recognition [1] [2], speech synthesis [3]. Audrey was the first speech recognition system developed in 1972 by Bell Labs. Speech Recognition has been widely used in number of activities like dictation tools in legal and medical profession, automated attendants in call centre, voice dialling.

2. **Music:** Acoustics has been used in field of music. Various musical applications like study of working of different musical instruments [4], different genres of music [5], human voice (singing), classification of audiovisual data into speech and music [6] employ acoustics. The branch of acoustics dealing with the music and music related applications is called Musical acoustics.

3. **Health Informatics:** In hospitals, acoustic modeling has a number of applications ranging from finding whether the patient has a disease or not to notifying the hospital staff of various events of patients. Doctors have been using stethoscopes to listen to the internal sounds of the human body since the early 19th century. Acoustical energy can be focussed and used for imaging and treating a variety of ailments including cancer, stroke, and Parkinson's disease. Video cameras can be another alternative for monitoring the patient activities but they have been criticized for intruding upon the patient's privacy. Acoustics have worked very well to discriminate between normal and pathological data [7].

4. **Aeroacoustics:** Acoustics can be applied to aircraft industry as it can help to analyze the sound of taking off and landing of aircraft. Also identification of various dangerous events like terrorist activities from the sound generated in aircraft is another application of acoustics in the air industry. Acoustic signals are also used in detecting any type of invisible damage in the panels of the aircraft which are not detected by staff inspection [8], detection of low flying aircraft [9], etc. Aeroacoustics help in studying the working of musical wind instruments.

5. **Underwater Acoustics:** Acoustics help in the prevention of accidents in water by detecting underwater objects such as icebergs. SONAR is an example of technology that uses underwater acoustics. It helps in detection of vessels underwater and reaches to them through sound. Tracking of vehicles under the sea, studying weather of the ocean, gathering information about marine life are some other applications of underwater acoustics.

6. **Acoustical surveillance:** Acoustics have been widely applied for security and safety, in short for urban surveillance. Various events such as sounds of closing or opening doors, dropping or breaking objects, gunshot, dog barking and screams can help in identification of various abnormal activities [10].

7. **Smart Homes:** Acoustics can help in building smart homes. Wang et al. [11] presented a robust environmental sound recognition system for home automation. Based on already classified sound sources, various services at home can be activated.

8. **Environmental noise and soundscapes:** Environmental acoustics is related to the noise and vibration of road traffic, recreational activities, railways etc. Acoustics can help to determine the context of the environment which can help in various context-aware applications.

1.2 Acoustic Classification Systems

There can be two types of Acoustic Classification Systems- Acoustic Scene Classification Systems and Acoustic Event Classification Systems.

1.2.1 Acoustic Scene Classification

Acoustic Scene Classification is defined as the recognition of the audio environment surrounding certain sounds. Based on social or physical context, the environment can be defined as a meeting room, office, healthcare, home, restaurant, park, traffic congestion etc [12] [13]. Acoustic scene classification assigns semantic labels to temporal regions of sound recording. Acoustic Scene Classification system is depicted in Figure 1.1. The system classifies the acoustic recordings into one of the categories- bus, office or park. An acoustic scene consists of multiple acoustic events.

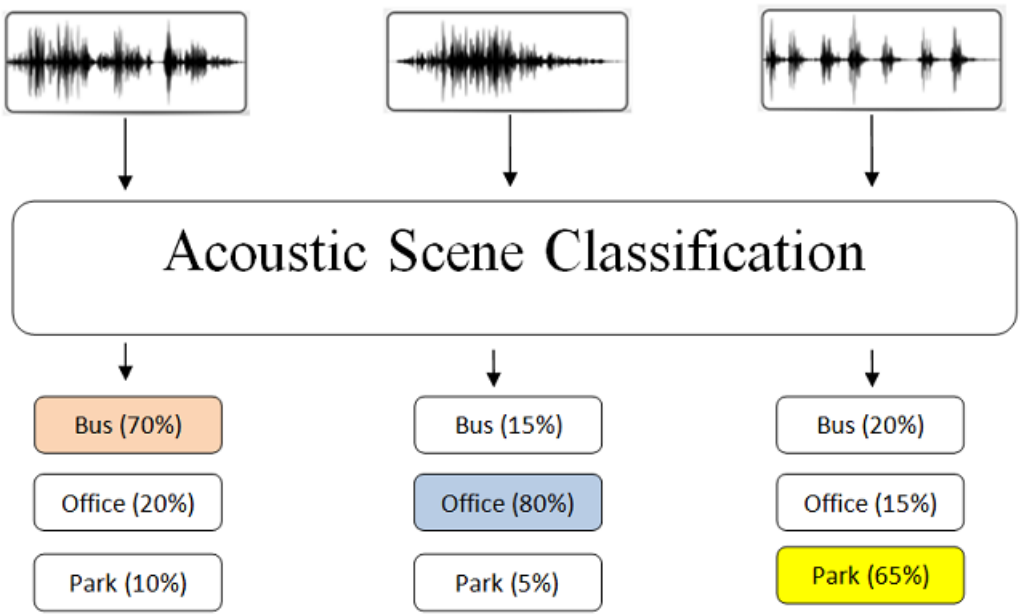

Figure 1.1: Acoustic Scene Classification

1.2.2 Acoustic Event Classification

An acoustic event is an occurrence of an audio signal at a particular point in time. Examples of acoustic events are bird singing, car passing by, sound of footsteps, closing or opening of door, person falling, gunshot, a group of persons talking etc [14]. An acoustic scene consists of multiple acoustic events. For example, the traffic congestion environment may have acoustic events like engine idling sound, honks, air turbulence, etc.

Acoustic event classification systems aim to classify sound events in the audio recognizing start and stop times of events.

1.3 Traffic event detection using acoustics

Monitoring road and transportation conditions and events related to them can help to provide important information. This information has several benefits for commuters and various other agencies. Various events may include traffic congestion detection, vehicle classification, and accident detection.

- **Traffic Congestion Detection:** Detection of the state of the road as congested or free-flowing can help in informing commuters about the traffic congestion state. Traffic congestion can be estimated from the speed of the vehicles [15] or from various events such as the number of the honks, air turbulence sound, engine idling sound etc [16].

- **Accident Detection:** Number of accidents are increasing day by day. The fatalities due to accidents are more because emergency responders do not reach early as they are not informed. Automatic accident detection can help in detecting road accidents and notify the hospitals and other concerned for help [17] [18].

- **Vehicle Classification:** Acoustics can help in classification of various types of vehicles like car, bus, bike [19]. Knowing the type of vehicle contributes to the surveillance applications. Automatic toll charge works on the principle of classification of vehicle type and imposing charges based on the type of vehicle.

1.4 Motivation

In today's era, the number of vehicles has increased enormously. Traffic congestion occurs when the number of vehicles increase more than the existing space of road. The problem of traffic congestion has already been addressed [20]. Though classification of road state into various congestion categories is important, little work has been done in the classification of vehicles on road.

Classification of vehicles on road means categorizing the vehicles on road under different labels. Classification of vehicles become more challenging on roads for developing countries like India where traffic is heterogeneous and chaotic. Vehicle Classification on roads can assist the traffic management policymakers, public safety organizations, insurance companies etc. Knowing the type and the number of vehicles plying on road, various

design decisions can be made by traffic management policymakers. Recognizing the type of vehicles aid in the number of applications like automatic toll collection, surveillance, accident prevention, traffic congestion avoidance, emissions/pollution estimation, traffic modeling etc [21].

There are various technologies already available that can help to find the class of vehicle. The existing technologies are either infrastructureless or infrastructure based. Infrastructure based technologies are based upon inductive loop detectors [22], magnetic sensors [21], video cameras [23] etc. Infrastructureless technologies are based upon GPS [24], Accelerometers [25] etc. But these have certain disadvantages as follows:

1. **High installation and maintenance costs:** The installation and maintenance costs of infrastructure based techniques are very high which poses the biggest disadvantage to the use of these technologies for traffic monitoring [20], [26]. Also, infrastructureless techniques such as GPS drain the battery very fast, if kept on.

2. **Occlusion and changed weather conditions (video cameras):** The accuracy of Vision based techniques is affected during night [27] and in different weather conditions such as rain, fog etc [28]. Further, during heavy traffic conditions, different vehicles may get occluded and it is difficult to detect vehicle.

3. **Non homogeneous and Non-lane driven traffic conditions:** Magnetic sensors and inductive loop detectors require the traffic to be homogeneous and move in the lane. In absence of that, their performance degrades [29].

Acoustics acquired from the vehicles can help in determining the type of vehicle. Microphones installed along the roadside can be used to capture acoustics from the road. They have low installation and maintenance cost [30]. As an alternative, smartphones can be used to capture the acoustics from roads [31]. Audio sensors have a number of advantages as compared to the above techniques:

1. Auditory-based system is not affected by occlusion and abnormal weather conditions etc.

2. Audio sensors are multidirectional, i.e., audio can be received from any direction.

3. Audio data is less sensitive to the location and orientation of the phone as compared with other common sensors such as cameras and accelerometers.

4. Though auditory-based traffic classification needs to perform preprocessing, normalization and typical feature extraction before giving the results but it is less computationally intensive as compared to camera-based or GPS-based traffic classification.

In this thesis, we have addressed the problem of vehicle classification using acoustics captured using commuters' smartphones. Smartphone based acoustic sensors are employed for collecting the sounds of vehicles such as vibrations in the engine, friction between the tires and the pavement, wind effects, gears, fans, sounds from rotational parts etc. These sounds are used for classifying vehicles on the roads.

1.5 Objectives

The main objective of our work is to determine the class of vehicle through sounds produced by them. To achieve this objective, we have undergone through the following study:

1. To collect, preprocess and extract various features from the acoustics collected using commuters' smartphones.
2. To analyze the efficacy of deep learning model for feature extraction.
3. To improve the vehicle classification accuracy using off the shelf CNN based approach as compared to human-defined features based approaches.

1.6 Organization Of Thesis

This thesis focuses on the new approach for vehicle classification through the sounds from vehicles, captured using the smartphones based acoustic sensors. Smartphones are the cost-effective way to acquire the sounds from the vehicles. Five generic classes of vehicles are identified for classifier:

- Car
- Bus
- Aeroplane
- Train
- Three-wheeler

The first part of this thesis outlines the domain of acoustics and its applications in various fields. It briefly explains the problem, motivation and the objectives for research.

Literature pertinent to vehicle classification on roads together with the technologies used by the researchers are surveyed in Chapter 2. It also introduces basic principles of acoustics, the classifiers employed in traffic related applications and research gaps.

The system overview and the methodology followed for the proposed approach is described in Chapter 3. It gives a detailed description of the approaches compared with the proposed approach.

Chapter 4 highlights the details of dataset used for the research, experimental setup and the results of the final proposed architecture in classifying the vehicles. The comparative performance analysis of the three approaches is also described in this chapter

Finally, the thesis concludes with the summary of the contributions, results of the proposed approach and the suggestions for future work in Chapter 5.

Chapter 2

Literature Survey

The analysis of the existing literature revealed that researchers have proposed several solutions for vehicle classification and other traffic related events detection. Intelligent Transportation Systems (ITS) monitor traffic conditions and also help in classification of vehicles plying on road. The data on the roads are collected through sensors. Several sensing techniques have been proposed which can be categorized into two categories-static sensor based techniques and movement based sensing techniques (Figure 2.1). The following two sections describe the static sensor based techniques and movement based sensing techniques respectively. The remaining sections briefly explain the basic principles of sounds, classifiers employed in traffic related events and architecture of Convolutional Neural Network(CNN).

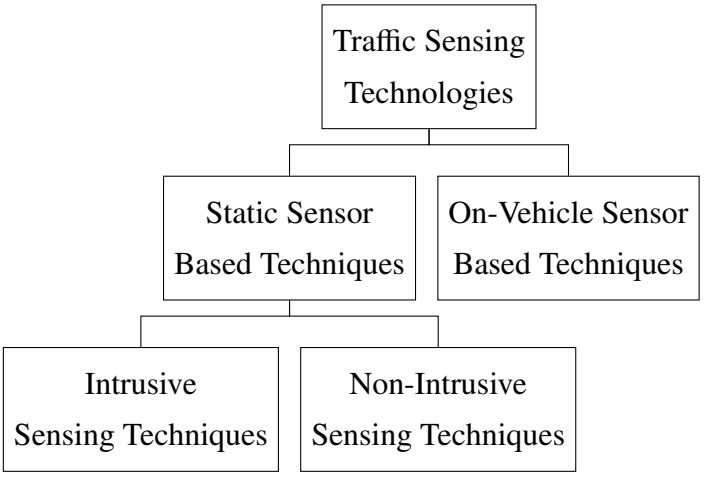

Figure 2.1: Traffic Sensing Technologies

2.1 Static Sensors Based Techniques

The techniques in which sensors are fixed are static sensors based techniques. These are either embedded under the road or installed along the road. Static sensors based techniques can be further classified as Intrusive Sensing Techniques and Non-Intrusive Sensing Techniques (Figure 2.1).

2.1.1 Intrusive Sensing Techniques

Intrusive sensors are installed on the road surface or pavement surface, in holes or in the saw-cuts of the surface of road, tunneled under the road surface [27]. Intrusive sensors include inductive loop detectors, magnetic sensors, pneumatic road tubes and piezoelectric sensors etc (Figure 2.2). Intrusive Sensing Techniques have certain drawbacks:

- Disruption of traffic while their installation and maintenance
- Failures in installation due to poor road surfaces and employment of substandard procedures for installation.
- Re-installation in case of repair and resurfacing of roads.

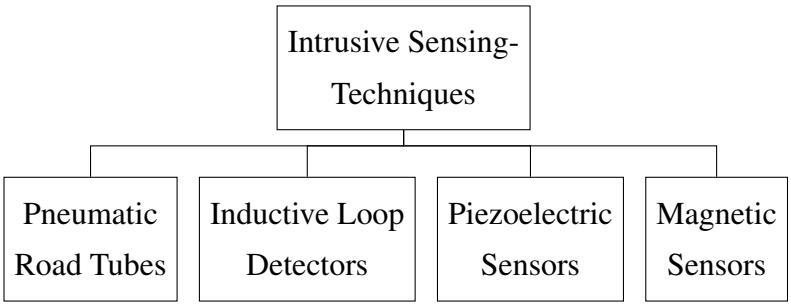

Figure 2.2: Intrusive Sensing Techniques

Pneumatic Road Tubes

The first commercially successful sensors for automatic vehicle detection were pneumatic road tubes. These are installed perpendicular to the direction of traffic flow [29] and help in detection of the vehicle. Classification, determination of speed and gaps between vehicles can also be determined if the number of tubes is increased [27].

Inductive Loop Detectors

Inductive loop detectors have been employed in vehicle detection and classification and hence help in traffic monitoring [29]. They work on the principle of electromagnetic induction (Figure 2.3). The inductive loop detectors can be employed for a number of applications like traffic state congestion detection [32], accident detection by determining the number of vehicles between two points [33] etc.

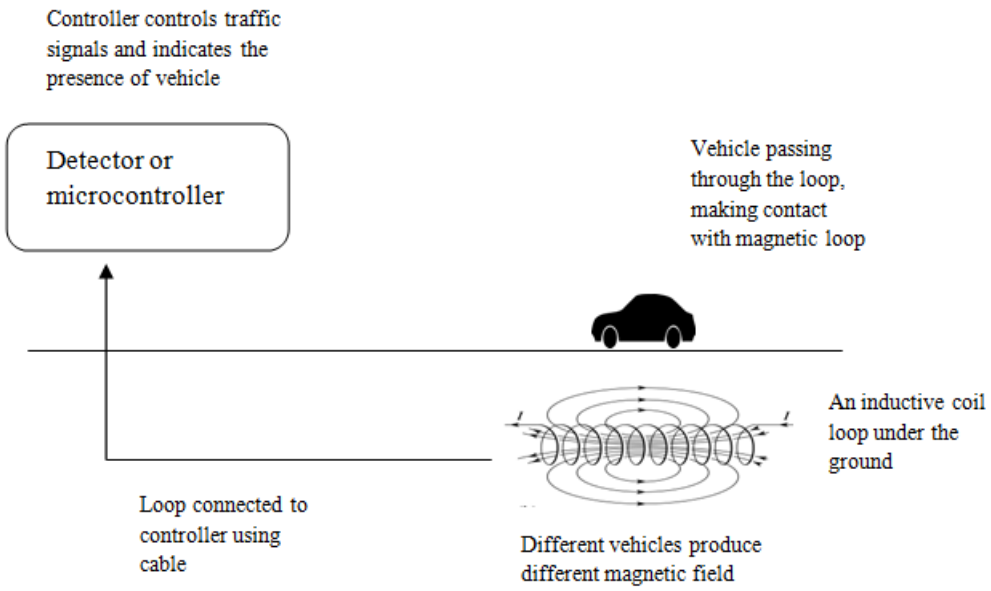

Figure 2.3: Inductive Loop Detector

Gajda et al. [22] explained the effect of loop length of inductive loop detector on magnetic profiles of vehicles that belong to different classes. By varying the length of the loop from 0.25-4 meters, they proved that shorter loop length can easily classify between different classes of vehicles. But this approach used the raw noisy data without preprocessing. So, Meta et al. [34] proposed a method in which noise of raw signal is first removed through Discrete Fourier Transform(DFT), then it was transferred to Principal Component Analysis(PCA) domain to decorrelate and reduce dimensions wihout losing any detail of signal. They used three-layer back propagation neural network(BPNN) to classify vehicles into five classes and obtained considerable accuracy of 94.21%.

Piezoelectric Sensors

Piezoelectric sensors measure changes in pressure, temperature etc. [27] and this information can be used for various applications like traffic monitoring [35], distinguing lanes [36] and measurement of various traffic parameters like speed, volume of vehicles [37] etc.

Sroka [38] presented the data fusion techniques by employing the parameters from piezoelectric sensor and the inductive loop detectors. Vehicles were classified into four classes with an accuracy of 92-94% but this combination was unable to discriminate motorcycles from other vehicles. So, a novel method based on single element piezoelectric sensors (Figure 2.4) was proposed which classified vehicles with an accuracy of 84.4% [39]. However, this method used average track width, so velocity and axle spacing were also not that accurate. Hence chances of misclassification of vehicles in other classes except motorcycles were very high. So, Rajab and Refai [40] proposed a single element piezoelectric sensor that used width to height ratio for vehicle classification. Though, the classification accuracy in this case was very high (98.9%-100%) but single element piezoelectric sensors are unable to give accurate track width. To overcome this limiation, multi-element piezoelectric sensor was used to classify vehicles into 13 classes accurately (86.9%-100%) based on the axle spacing [41].

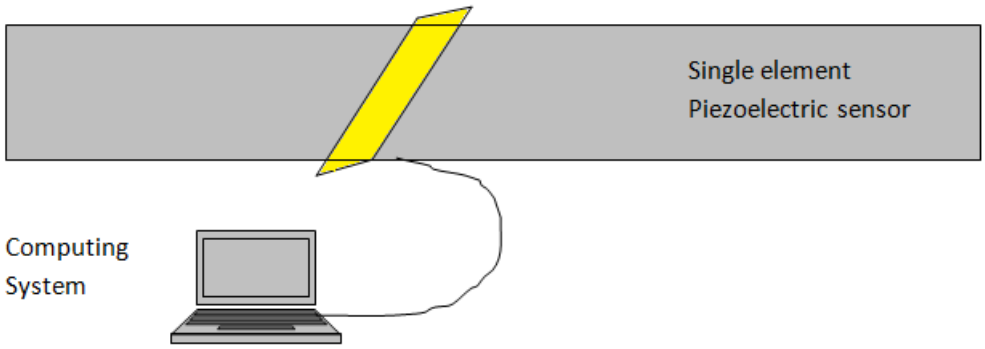

Figure 2.4: Single Element Piezoelectric Sensor

Magnetic Sensors

Magnetic sensors are passive devices that detect the presence of the vehicle by measuring the change in magnetic field of the earth which occurs when the metallic vehicle is passed over the sensor [29]. De Angelis et al. [21] proposed a novel method based on magnetic signatures for vehicle detection and classification in smart cities. They estimated length and speed of vehicles, hence the class of vehicles and reported a high accuracy of 98%. But the magnetic sensors were to be re-installed when repairing of roads was done (Table 2.2). So, Kaewkamnerd et al. [42] proposed wireless magnetic sensors for automatic vehicle classification. Vehicles were classified into four classes and acceptable accuracy was obtained- car (68.42%), pickup (82.86%), van (71.43%) and motorcycle (95.83%).

2.1.2 Non-Intrusive Sensing Techniques

Non-intrusive sensors are mounted on the sides of the road or above lane of the road at a certain angle to the direction of traffic [28]. The road traffic is not disrupted while their installation and maintenance and have the same accuracy as intrusive techniques. These technologies include microwave radar sensors, acoustic sensors, infrared sensors and video cameras (Figure 2.5). They help to find vehicle presence, count, passage, class, speed and multiple-lane detection zone coverage [20].

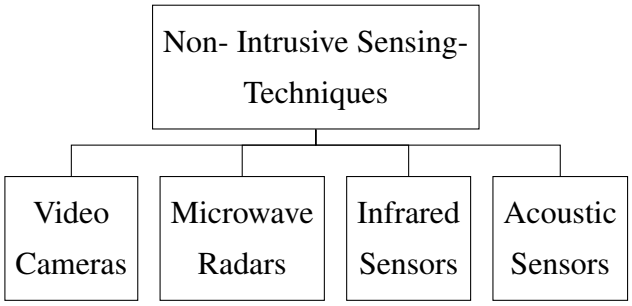

Figure 2.5: Non-Intrusive Sensing Techniques

Video Cameras

Video cameras are extensively used for calculation of traffic parameters and hence in traffic monitoring. The deployment of video cameras at some height capture the state and types of vehicles plying on roads. They help in detection of vehicles [43] [44] [45], their speed [46], class, average speed, number of vehicles per unit time [47] and also help to detect accidents [48].

Hasegawa et al. [23] described a vision-based system that provided for vehicle detection and vehicle classification on roads through video sequences. They claimed that classification accuracy of 91.1% is obtained. But they failed to handle the occlusion in video sequences. So, a technique to handle partial and full occlusions was proposed [49].The computer vision technology was employed to classify vehicles into four categories- motorcycle, car, lorry and background (without vehicles) [50]. Though, high classification accuracy (92.3%) was obtained but it was not robust to variations such as illumination, weather, noise etc. So, Zhuo et al. [51] proposed the new approach using video sequences and Convolutional Neural Network for vehicle classification.

Microwave radar

Microwave radar transmits low energy microwave radiation with the frequency in range 2.5 to 24 GHz into the zone in which vehicles are to be detected [28]. Vehicles reflect back the signal which is analyzed for monitoring traffic [52], determining the class of vehicles or tracking the trajectory of vehicles and collision avoidance [53]. Cherniakov et al. [54] employed CW Doppler radar system (Forward scattering radar) to collect vehicle signatures and to classify vehicles into two classes-tracked and wheeled with an accuracy of 79.7%. They could not classify vehicles into more number of classes. So, a vehicle classification system that classified vehicles into five categories with considerable accuracy of 99% was developed [55].

Infrared Sensors

Infrared sensors are installed overhead to view approaching and departing vehicles [28]. They help in various traffic related applications such as detection of traffic state congestion [56], determination of the volume of vehicles and speed of vehicles on roads [27] (Table 2.2). For the first time, the usage of infrared sensors for a very large data volume was presented by Tropartz et al. [57]. They collected data of 2.3 million vehicles from 8 commercial toll plazas, out of which 12.8% was heavy traffic and this approach yielded an accuracy of 98.5%. Odat et al. [58] developed a system based on passive infrared sensors which helped in traffic monitoring, vehicle classification and speed estimation of vehicles. The authors claimed that the accuracy of 99% was obtained.

Acoustic Sensors

Traffic state can be monitored using acoustic signals captured by acoustic sensors. Various traffic parameters like class, presence, the passage of vehicles can be determined through sounds produced by vehicles on roads [28]. The presence of the vehicle is detected by setting some detection threshold. The increase in sound energy above detection threshold occurs when the vehicle enters the area of detection and reverse occurs when the vehicle exits the area of detection [27].

Many researchers have developed systems that use the acoustics to monitor traffic parameters. Maciejewski et al. [59] proposed a neural network based approach for classifying military vehicles into four classes based on sounds produced by them. First, preprocessing was performed using wavelet multiresolution analysis and the feature vectors extracted were used for classification of vehicles. The classifiers- Probability Neural network(PNN) based on Gaussian mixture and Multilayer Perceptron(MLP) were used

band it was asserted that the error rate in case of MLP is slightly higher than PNN. 'Eigen faces Method' was used to discriminate cars from other vehicles [60]. The frequency spectra of 200 ms obtained from each audio signal formed the part of training set. The sounds were characterized using two vectors- Mean vector of the whole training set and eigenvector. During testing, the frequency spectrum of the test vector was subtracted from the mean vector, principal component analysis was performed over the difference vector and residual was found. This unknown vector's coefficients gave the class of vehicle.

Though the researchers have succeeded in obtaining high accuracy in vehicle classification but the classes to which vehicles were categorized were small. So, Munich [61] proposed an approach for classifying vehicles into mainly 9 classes using Bayesian subspace method. The effect of frame size was analyzed and the efficacy of three classifiers- Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) and Bayesian Subspace method was compared. Bayesian subspace method gave 50% higher accuracy (error rate of 11.7%) as compared to GMM (error rate of 24.85%) and HMM (error rate of 23.10%). It was proved that smaller frame size of 256 is better than larger frame sizes like 512.

Sen, Raman, et al. [62] proposed an inexpensive technique to estimate the speed of the vehicles and to determine the traffic state of the road using acoustics. They deployed two acoustic sensors and doppler frequency shift principle was used to estimate the speed of the vehicle. Sequence of algorithms were employed to detect honk, match the honk between two sensors and extract the frequency from the honk. The estimated speed classified the traffic state of road as congested or free-flowing with an accuracy of 70-100%. This was practically deployed as hardware to handle chaotic traffic [16]. The hardware was deployed on six roads for six days and Support Vector Machine (SVM) classified the traffic state with an accuracy of 92.7-100%. However, the large-scale deployment of hardware and changing batteries every 2 days were cumbersome and expensive.

To overcome the above limitations of hardware, Tyagi et al. [20] proposed the use of cumulative acoustic signal captured through roadside installed omnidirectional microphone to estimate the density of vehicular traffic. They conducted experiments to determine the best window size (40, 100, 200, 500 ms), shift size (20, 50, 100ms) and signal duration. Bayes' classifier and SVM classifier gave the accuracy of 95% and 100% respectively with the window size of 500ms, shift size of 100ms and time duration of 30s.

Vehicle detection and classification were performed using Artificial Neural Network [63]. In this method, log energy features were given as input to Artificial Neural Network and vehicles were classified into four categories- horns, medium, light and heavy vehicles with an accuracy of 66-67%. The raw audio signal without preprocessing were used. So, George, Mary, et al. [64] proposed a new technique in which the sounds collected through

microphones are preprocessed to remove noise. Artificial Neural Network(ANN) and K-Nearest Neighbour(KNN) classified the vehicles into three categories- heavy medium and light with the accuracy of 73.42% and 50.62% respectively. KNN gave the accuracy of 50.62% and ANN of 73.42%.

Paulraj et al. [19] aimed to find the class and position of a vehicle using the acoustics collected from the vehicles. Experiments were conducted to determine the best number of consecutive frames (1,2,3,4,5) and the size of training set (60%,70% and 80%) that can yield the highest accuracy. Probability Neural Network gave the best results for 4 consecutive frames and training set size of 80%. They stated that the accuracies of 94.5% and 92.8% were obtained for classification of vehicles' types and vehicles' positions respectively. The vehicles were not categorized within the classes. This was attempted by Kandpalet al. [65] who classified the ground vehicles on road and identified vehicles within the class. Multilayer Neural Network gave an accuracy of 80% for classification of ground vehicles. For classification of the vehicle within the individual classes, accuracy obtained was 57% in case of car and truck, 60% in case of the bike.

A few researchers have employed acoustic sensors present in smartphones to collect data from the vehicles and employed that for monitoring traffic conditions. Kaur et al. [31] proposed the use of smartphones based acoustic sensors for capturing sounds from roads and determined the traffic congestion state. The authors experimented with different frame sizes (1024, 2048, 4096, 8192ms), window functions, shift sizes (50, 60, 70, 80, 90%), feature sets (RMS, ZCR, STE, MFCC, Delta and Delta-Delta) and classifiers (Neural Networks, SVM) to determine the best combination of parameters. SVM gave better accuracy (93%) as compared to Neural Networks (91.8%) for frame size of 8192ms, hamming window function and feature set of MFCC, STE and RMS. The change of shift size had no effect on the classification accuracy.

The work done by various researchers for determination of traffic parameters using acoustics is illustrated in Table 2.1.

Table 2.1: Determination of Traffic Parameters Using Acoustics

Author/ Year	**Database**	**Classifier and Performance**
Maciejewski et al. 1997 [59]	Self collected in different road conditions.	Probabilistic Neural Network based on GMM(PNN) and Multilayer Perceptron (MLP). Error rate of MLP for classification of military vehicles is greater than PNN by approx. 2.5%.

Author/ Year	Database	Classifier and Performance
Wu et. al. 1998 [60]	Self Collected	Eigen Faces Method performed fairly good for vehicle detection and discrimination of cars from other vehicles .
Munich 2004 [61]	ACIDS database.	Bayesian subspace gave 50% higher accuracy than GMM and HMM for vehicle classification into 9 classes.
Sen, Raman, et. al. 2010 [62]	Self Collected: 18 hours	Three algorithms based on Doppler Principle for traffic state detection and accuracy of 75-100%.
Sen, Siriah, et. al. 2011 [16]	Self Collected	Support Vector Machine (92.7-100%) for vehicle detection and classification.
Tyagi et. al. 2012 [20]	Self Collected	Bayes' classifier (95%) and SVM (100%) for traffic congestion detection. Window size of 500ms, Shift size of 100ms, Signal duration of 30s.
Paulraj et al. 2013 [19]	Self Collected. Three training datasets (60, 70,80% of data)	Probability Neural Network (PNN). Vehicle classification (90.6, 93.9, 94.5%)and Vehicle Position(89.8, 90.5, 92.8%) for 60%, 70% and 80% data respectively.
Kandpal et. al. 2013 [65]	Self Collected. Three classes: car, truck and bike.	Multilayer Neural network classification approach with an accuracy of 80% and an accuracy of 57% for car and truck, 60% for bike within the class.
George, Cyril, et al. 2013 [63]	Self collected. Each recording of 15-30 mins.	Artificial neural network with an accuracy of 66-67% for vehicle detection and vehicle classification.
George, Mary, et. al. 2013 [64]	Self collected. 160 vehicles' recordings of different categories.	ANN and KNN with an accuracy of 73.42% and 50.62% respectively for vehicle detection and vehicle classification..

Author/ Year	**Database**	**Classifier and Performance**
Kaur et al.2017 [31]	Self collected using smartphones: 320 recordings	SVM (93%) and Neural Networks (91.8%) for traffic state detection.

2.2 On-Vehicle Sensor Based Techniques

The sensing techniques in which sensors are neither buried under the roads or pavements, nor they are installed on the roadside, those are called On-Vehicle Sensor Based Techniques. These sensors are deployed either in smartphones or probe-based vehicles so, they can be carried everywhere. These techniques are called off-road sensing techniques. Two on-vehicle sensor based techniques- Global positioning system(GPS) [66] and Accelerometers [67] are discussed below (Figure 2.6).

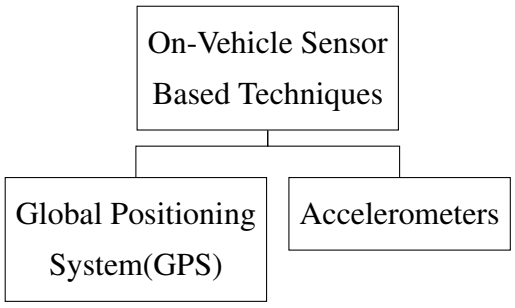

Figure 2.6: Movement Based Sensing Techniques

2.2.1 Global Positioning System (GPS)

GPS is a satellite-based navigation system [66]. GPS works 24/7 and in all weather conditions. GPS is widely used for calculation of traffic parameters [68]. They can be installed on smartphones or vehicles for determining the vehicle parameters.

GPS in vehicles

Nowadays, GPS is installed in the number of public transports like buses and cars to track them. Vehicles' speed and location information can be disseminated to Traffic Management Centre(TMC). They provide the advantage of finding out the travel time and speed of vehicles [69], traffic congestion state [70] and vehicle classification [71].

Smartphones with GPS

GPS present in smartphones transmit the data collected to control center. This in-built equipment of mobile phones with GPS is cost-effective as unnecessary hardware need not be installed, unlike GPS in vehicles. Many researchers have exploited the GPS present in the smartphones to gather the information about traffic parameters like traffic congestion state [72], bumps and potholes detection [73] and vehicle classification [24].

2.2.2 Accelerometer Sensors

Accelerometers in smartphones can be used for monitoring traffic by detecting the change in orientation. This helps in detecting traffic parameters and surface of roads like potholes and speed breakers [74], determining traffic congestion state [75], determining road conditions and driving conditions [76], classifying vehicles on roads [25].

The static and movement based sensing techniques are summarized in Table 2.2

Table 2.2: Sensing Technologies for determining Traffic parameters

Technologies	**Capabilities**	**Strengths**	**Weakness**
Pneumatic Road Tubes [29]	- Detection of vehicles, determination of speed and gap between the vehicles. - Classification of vehicles	-Low power usage and low cost. -Easy and simple to install and maintain	-Cannot detect stationary vehicles over sensors. -Cannot discriminate between lanes. -Temperature-sensitive
Inductive Loop Detectors [34] [22]	-Detection, speed, count of vehicles. -Classification of vehicles:Different vehicles produce different magnetic field	-Resistant to abnormal weather like rain, snow etc. -Determines common parameters of traffic like volume, speed, occupancy, gap.	-Temperature and traffic sensitive. -Buried under the roads so require pavement cut. -High installation and maintenance for high scale monitoring. -Multiple loop detectors for monitoring.

Technologies	**Capabilities**	**Strengths**	**Weakness**
Piezoelectric sensors [38] [40]	-Determines vehicle count and speed. -Classification of vehicles:Different vehicles exert different pressure.	-Less costly and more accurate than inductive loop detectors. -Can monitor up to four lanes. -Can be used for high speed. vehicles(16 to 112kph)	-Temperature and traffic sensitive. -Jamming of roads while installation and maintenance. -Repair and resurfacing of roads require sensors' re-installation.
Magnetic Sensors [21] [42]	-Detection of vehicles. -Determination of speed,number and direction of vehicles. -Classification of vehicles:Magnetic field changes with changing vehicle.	-Less costly as compared to inductive loop detectors. -Effective where loops are not feasible(bridge decks). -Less sensitive to traffic stress.	-Can not detect stationary vehicles. -Small detection zones. -Work well for lane driven and homogeneous traffic. -Sensitive to temperature and noise change.
Video cameras [23] [51]	-Determines presence, speed, count and occupancy of vehicle. -Classify vehicles taking into account shape.	-Ability to monitor multiple zones. -Provide rich data. -Gather information from several locations and detect wider area.	-Non-resistant to abnormal weather conditions like rain and storm. -Affected by occlusion. -Can not provide 360 degree view.
Microwave Radar [54] [55]	-Determine presence(FMCW radar) direction, speed and number of vehicles. -Classify vehicles. -Track the trajectory of vehicles.	-Unaffected by adverse weather conditions. -Can detect multiple zones. -Can operate both during day and night.	-CW Doppler radar can not detect stationary vehicles. -Work poorly at intersections as traffic increases.

Technologies	Capabilities	Strengths	Weakness
Infrared Sensors [57] [58]	-Vehicle presence, speed, count and direction. -Classification: Depending upon shape	-Detect traffic of multiple lanes. -Need not be installed in pavements. -Can work in day as well as night.	-Gleam of sunlight cause unwanted signals. -Affected by rain, snow etc. -Affect normal traffic while installation and maintenance.
Acoustic sensors [19] [31]	-Determines the speed, count, volume, direction and presence of vehicles. -Classification of vehicles.	-Less costly than other sensors. -Unaffected by precipitation. -Efficient in chaotic and heterogeneous traffic.	-Inaccurate in case of cold temperatures. -Some of the acoustic sensors do not work in stop and go type of traffic.
Probe based vehicles [71] [25]	-Determine vehicle speed,direction and volume.	-Less costly. -Fixed point installation is not required on the road.	-Vehicles installed with sensors are compulsory.

2.3 Basics of Acoustic Processing

Audio data from the vehicles on the roads can be acquired by the audio sensors. The data is processed to determine the traffic parameters such as the class, speed of vehicle etc. Basic principles of acoustics, varieties of features, area of application of features and varieties of classifiers that are used in audio classification are discussed below.

2.3.1 Basic Principles of Sound

Sound travels through the air as the continuous wave of pressure changes. It is produced due to vibrations. Sound has the properties of refraction, reflection, diffraction etc. There are a variety of sources that generated sounds. Direct pressure variations (Acoustic), air pressure changes and loudspeaker like things produce sound. On the other hand, ears and electrical devices(microphones) receive sounds which convert pressure changes to electrical signals. The electrical devices like microphone produce the sound

that is analog in nature. If the sound is to be processed or stored in the computer, it must be in digital format(stream of numbers). So, the analog signal needs to be digitized. This process is named digitization of sound. The analog signal of sound can be represented as a waveform in which the horizontal axis represents time in seconds and the vertical axis represents signal strength in volts.

Digitization of Sound

Analog to Digital converter is employed to convert the analog signal to the stream of bits. Digitization of sound is carried out in two steps: Sampling and Quantization. Before discussing sampling and quantization, we need to know certain parameters that determine the quality and amount of information in digitized sound. Parameters are given below:

- **Sampling rate:** Sampling rate is the number of samples of the sound signal taken per second. The unit of sampling rate is Hz. The range of sampling rate in sound is between 8kHz to 44 KHz. The sampling rate is determined using Nyquist Theorem. This is also called Nyquist sampling rate according to which the sampling rate is twice the highest frequency in the input signal.

- **Bits per sample:** The amount of information in each sample is called Bitrate. The amount of information is determined by the number of bits of resolution of each sample. More bits per sample require more bandwidth and the quality of the audio is very high.

- **Mono v/s Stereo:** In Mono or monaural sounds, a single channel is there through which audio is routed. In stereo or stereophonic sound, two channels are there through which audio is routed. The size and bandwidth required by mono are less than stereo. Stereo sounds are used to create multitrack recordings and provide direction and location information of sound.

The digitization process is carried in two steps:

1. **Sampling:** Sampling is digitization's first step. The horizontal axis (time axis) is divided into equally spaced intervals. It involves the determination of the amplitude of air pressure at equally spaced locations along the time axis. Measurement of amplitude at each location is called sample. Sampling is a reversible process.

2. **Quantization:** The second step of digitization is Quantization. It discretizes the vertical axis (signal strength) into different levels. A fixed bit number can be used to represent each level.This is a lossy process and is irreversible, unlike sampling.

Framing

Since an audio signal is a continuous waveform, statistical properties of the signal keep on changing with time. So ,the signal is splitted into frames such that each frame is considered pseudo stationary [31]. This splitting of the frames is needed for short-term analysis. There can be several techniques used for framing-sliding window, activity defined window and event defined window. Sliding window technique has attained wide popularity in real-time applications because it is simple and does not require any pre-processing.

Depending on the event length and type, frame size is decided. Smaller frame size means more number of frames as the signal will be split into many chunks. Larger frame size leads to capturing of multiple events. Also, the processing of larger frames takes more time. The number of frames can be calculated if sampling rate and frame size are known.

$$N = t * fs \tag{2.1}$$

Where N is the number of frames,t is size of frame and fs is sampling rate.

There can be overlapping among the frames. Overlapping is advantageous as the events happening at the discontinuity are included in the frame that is overlapped, otherwise, the events at discontinuity are lost. Sometimes there is valuable information in lost events (Kaur, Sood, Aggarwal, *et al.* [31]). Also, if training data is less or more frames are required, overlapping among the frames helps.

Windowing

Windowing eliminates the sharp discontinuities of the frames. It is considered as the smoothening function that is applied over the frames. The amplitude of the side lobes and edges is highly attenuated. Windowing functions are zero outside some chosen intervals. If some other data sequence or waveform is multiplied with the windowing function, it also becomes zero-valued outside the interval. Only the part which overlaps with the window is left. There are varieties of window functions that can be applied. They are described below:

1. **Rectangular Window:** The most simple window is the rectangular window (Figure 2.7). It is constant inside the interval and zeroes outside the chosen interval. The effect of applying the rectangular window is equivalent to framing with the sliding window.

$$w(m) = \begin{cases} 1, & m = 1.....N-1 \\ 0, & \text{elsewhere} \end{cases}$$

This window is good for inspecting those signals which have very small duration, called transients. For e.g. a noise burst, an impulse, a shock signal etc. High spectral leakage is caused due to the rectangular window.It performs poorly for sinusoids having disparate amplitudes.

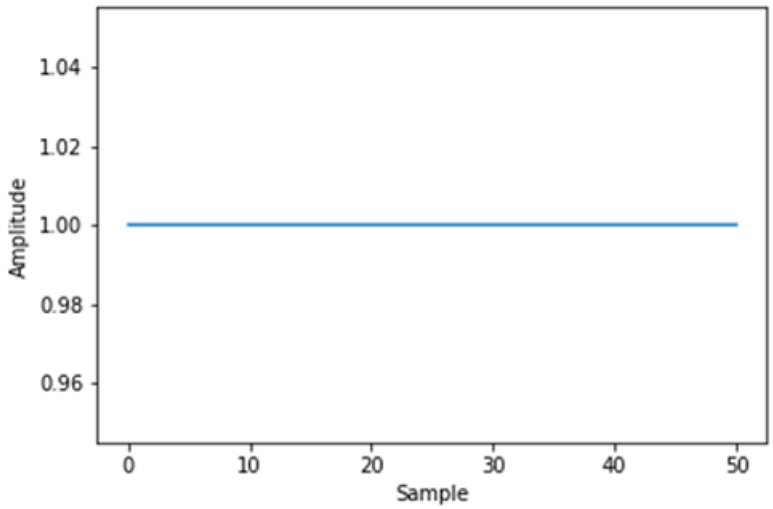

Figure 2.7: Rectangular Window

2. **Hanning Window:** Julius von Hann proposed Hanning windows. It has a sinusoidal shape (Figure 2.8). It results in low side lobes and wide peak. This window touches zero at both the ends i.e. side lobes hence, all discontinuity is eliminated. Due to this property, Hanning window is used for inspecting the transients having length longer than the window size.

$$w(m) = 0.5\left(1 - cos\left(\frac{2\pi m}{N-1}\right)\right) \tag{2.2}$$

3. **Hamming Window:** Richard W. Hamming proposed Hamming window. The size of hamming window and frame is kept same. This window also has a sinusoidal shape. It also results in low side lobs and wide peak but this window does not touch zero at both the side lobs, so there is a minor discontinuity in the waveform. So only nearest side lobs are canceled by the window (Figure 2.9). Harmonics of audio are improved by the hamming window.

$$w(m) = \alpha - \beta cos\left(\frac{2\pi m}{N-1}\right) \tag{2.3}$$

where α =0.54 and β =1-α =0.46

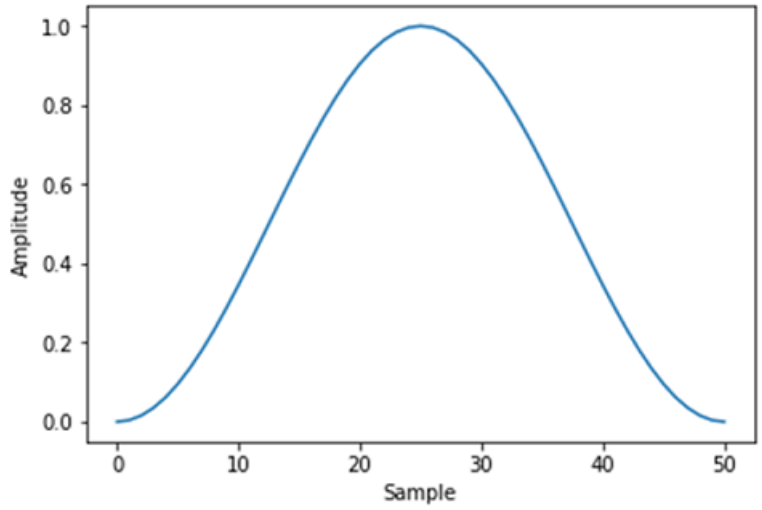

Figure 2.8: Hanning Window

$$w(m) = 0.5\left(1 - cos\left(\frac{2\pi m}{N-1}\right)\right) \tag{2.4}$$

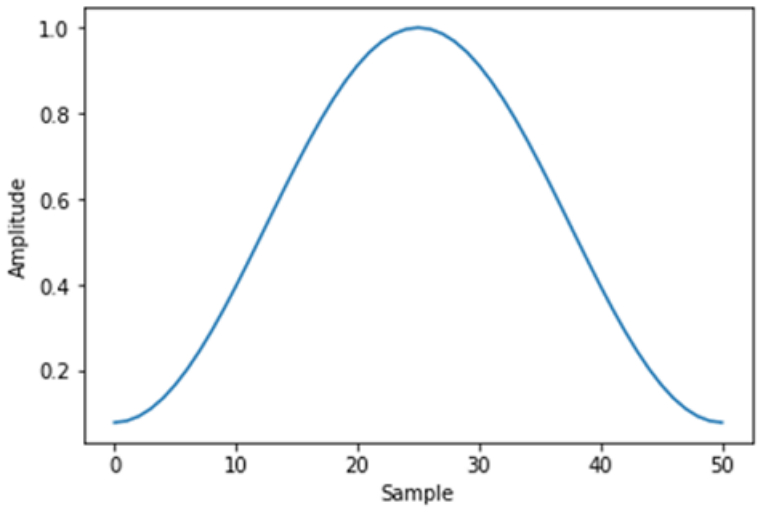

Figure 2.9: Hamming Window

4. **Blackman Window:** Blackman window is much the same as Hamming and Hanning windows. When Blackman window is applied over the signal, the spectrum formed has the broad peak but side lobe compression is good (Figure 2.10). The equation for Blackman window is:

$$w(m) = a_0 - a_1 cos\left(\frac{2\pi m}{N-1}\right) + a_2 cos\left(\frac{4\pi m}{N-1}\right) \tag{2.5}$$

where $a_0 = \frac{1-\alpha}{2}; a_1 = \frac{1}{2}; a_2 = \frac{\alpha}{2}$

$$w(m) = 0.5\left(1 - cos\left(\frac{2\pi m}{N-1}\right)\right) \tag{2.6}$$

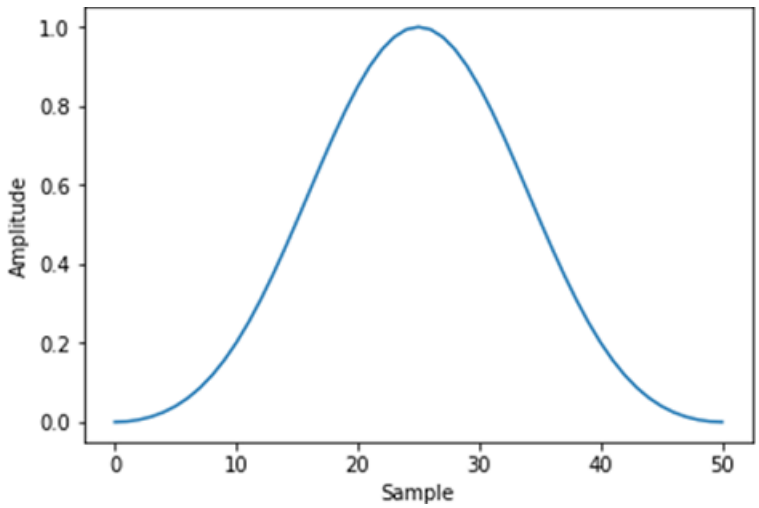

Figure 2.10: Blackman Window

Windowing Parameters

Depending on the kind of application to be developed, various windowing parameters are selected. Parameters such as overlapping size, window type, and frame size are considered.

- **Window Type**

  Selection of the appropriate window function is important. When the signals are shorter than windows (transients), then rectangular window is used. Since it is constant inside and zero outside, so it will generate false results in case of larger signals as the signal will be cut off at the boundaries. The most widely used window is the hamming window as it reduces the signal towards zero at boundaries and hence helps in reduction of discontinuities.

- **Frame Size**

  It is very important to consider appropriate size of frames. Very small frame size is needed to consider each frame as pseudo-stationary as the signal is continuously changing over time. Framing is necessary to capture the details of the rapidly changing signal.

Fourier transform generates frequency vector which is the combination of frequency bins. The width of frequency bins is affected by the number of frames analyzed and sampling rate.

$$Width = \frac{FS}{N} \tag{2.7}$$

where FS is sampling rate and N is the number of frames. So the number of frames or signals have an inverse relation with the width of frequency bin. It means larger frame size is required to get high frequency resolution.

Thus, there is trade-off in deciding the frame size. If the larger frame size is considered then the signal will keep on varying and it can not be considered stationary but the frequency resolution will be very high. On the other hand, if the smaller frame size is considered then frequency resolution will be less but the signal can be considered as stationary. So depending on the problem being considered, frame size has to be decided. Audio signals that contain traffic state should have a larger size of frames as the state of traffic does not change very rapidly [20] [31].

- **Overlapping Size**

  Overlapping in frames is required to capture the events that occur at the discontinuity. The size of overlapping or in contrary shift size need to be decided. Shift size is inversely proportional to overlapping size. Overlapping size is dependent upon the kind of application. For traffic state determination by Tyagi et al. [20], shift size (20ms, 50ms, and 100ms) was varied to determine the best shift size. The shift size of 100ms gave the highest accuracy. But in Kaur et al. [31], the different shift sizes gave not much difference in accuracy. Larger overlapping sizes give smoother results but the computational expense of the system is increased.

2.3.2 Feature Extraction

The process of extraction of characteristics of the sound is called feature extraction. The characteristics extracted are called features and they can be redundant or informative. It is necessary to extract informative features as bad features can make the problem hard to solve. Good features, on the other hand, will make the problem easy to solve and will provide a lot of important information. There are the number of feature extraction techniques and various types of features can be categorized into two main categories: temporal features and spectral features.

There are variety of features but all the features are not required in all applications. Hence, all the features need not be extracted. So, depending upon the application, different combination of features should be selected. The main objective of the proposed

approach is to classify the vehicles on the road using smartphone-based acoustic sensing. Since data collection is through smartphones, so various factors like classification accuracy, power consumption, and computational complexity need to be considered while deciding the different combination of features to be used [31]. To select the adequate combination of features for any application, following factors should be considered:

- The feature vector selected should be of the small size as more computation is required to process larger feature vectors.
- The classification accuracy of feature vector selected should be high so that they can differentiate easily between two classes.
- The technique selected for feature extraction should have low computational cost and time.

Temporal Features

Temporal features are also called time domain features. They are drawn out directly from the sound signal and no pre transformation is required. Hence, the computational complexity is low. The audio signal has the following temporal features.

1. **Zero Crossing Rate (ZCR)**

   ZCR is the frequency of sign changes of the signal. In other words, it is the frequency with which signal crosses zero that is the sign change from negative to positive or vice versa [31]. Banchhor et al. [77] worked to classify different musical instruments-guitar, harmonium, and flute by calculating zero crossing rate. Panagiotakis et al. [6] developed a method of classification of audiovisual data into speech and music. ZCR has an advantage of fast computation as spectrum computation is not required for it. But it is affected by noise.

   $$ZCR = \frac{1}{2N}(\sum_{n-1}^{N} |sgn(x(n)) - sgn(x(n-1))|) \tag{2.8}$$

   where x is a time domain signal, N is the size of processing frame and sgn is signum function. The signum function is given by:

   $$sgn(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases} \tag{2.9}$$

2. **Short Time Energy (STE)**

   The energy of audio changes with time. The energy of short-term region of audio data is called short time energy. It measures how the amplitude varies over time. It is also used to distinguish voiced speech, unvoiced speech and silence from each other. The amplitude of voiced speech is higher than the unvoiced speech. Silence has no energy. Short-term energy is a square function, so it is affected by large signal amplitude levels.

$$STE = \frac{1}{N}\sum_{n=1}^{N}(x(n))^2 \tag{2.10}$$

   where $x(n)$ is a time domain signal and N is frame size.

   Li et al. [78] used short-term energy for classification of audio data into seven categories. Then Lu et al. [79] classified the audio data into four classes in two steps using variation of short term energy. ZCR and STE together are used for classification of different musical instruments [77] since they have low computational complexity and can be easily computed.

3. **Autocorrelation**

   Autocorrelation is used to measure the similarity between a signal's value at a certain time point and its value when it is shifted by some time. It is used to find periodic patterns in the signal which are made indistinguishable by noise. This is done by finding at what time lag the signal's value repeats. The value of the signal is multiplied by its value at certain time lag and it is repeated for all possible time lags in the signal. At last, the values from series of multiplication are added and an average is computed. Autocorrelation is used to distinguish harmonic and non-harmonic sounds.

$$AC = f_{xx}[\tau] = x[\tau] * x[-\tau] = \sum_{N-1}^{n=0} x(n).x(n+\tau) \tag{2.11}$$

   where τ is a lag in the signal, $f_{xx}[\tau]$ is the value of autocorrelation, N is frame size. If τ is zero, then $f_{xx}[\tau]$ becomes the signal's power. Zero-crossing rate, short time energy and autocorrelation are used to distinguish voiced and unvoiced speech signals [80]. Voiced segments have high energy and low ZCR and remain periodic after application of autocorrelation function. The reverse is true for unvoiced segments.

4. **Root Mean Square (RMS)**

   RMS, also called quadratic mean is one of the features to measure the energy of the signal. First the arithmetic average of squares of sound signals is computed and then the square root is taken. The resultant is the value of Root Mean Square error. It is based on the principle that higher weights should be given to larger values and it is always greater than or equal to the arithmetic mean.

$$RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x(n))^2} \tag{2.12}$$

   where $x(n)$ is a value of the signal,N is frame size.

   RMS and ZCR can be used to distinguish speech and music [6].

5. **Energy Entropy (EE)**

   Energy Entropy is a parameter that computes and gives an indication of instantaneous change in energy level of the signal. Each frame is split into sub-frames and the energy of each subframe is calculated as in equation 2.10. The energy of each subframe is divided by energy of the frame.

$$e_j = \frac{E_{subFrame_j}}{E_{shortframe_i}} \tag{2.13}$$

   where

$$E_{shortFrame_i} = \sum_{k=1}^{K} E_{subFrame_k} \tag{2.14}$$

   Finally the energy-entropy is calculated as:

$$H(i) = -\sum_{j=1}^{K} e_j.log_2(e_j). \tag{2.15}$$

   Energy entropy is a useful feature in audio scene characterization. Violent scenes are characterized by determining abrupt changes in energy [81] [82] [83]. If a variation in energy is large in some frame then, the energy-entropy of that frame is less.

Spectral Features

Spectral features are derived from temporal features by exposing the temporal features to some transformation. Temporal features are converted into frequency based features called spectral features. The transformations for conversion can be Discrete Wavelet Transform, Discrete Cosine Transform or Discrete Fourier Transform. Discrete Fourier

Transform has low computational complexity, so it has a widespread implementation. Also, sometimes the banks of bandpass filters are used to transform temporal to the frequency domain e.g.Mel filters. Common Spectral features for audio signal are spectral entropy, spectral centroid, spectral roll-off, spectral flux, Mel frequency Cepstral Coefficients(MFCCs) etc. Spectral Features are discussed below:

1. **Spectral Entropy**

   Spectral entropy is computed in the frequency domain, which is the only thing that makes it different than the entropy of energy discussed above. The short-term frame is divided into L subframes. The energy of each subframe(let be f) is computed and then it is normalized by dividing it by overall frame's energy.

   $$e_j = \frac{E_{subFrame_j}}{E_{shortframe_i}} \tag{2.16}$$

   where

   $$E_{shortFrame_i} = \sum_{k=1}^{K} E_{subFrame_k} \tag{2.17}$$

   and then entropy is calculated using formula written below:

   $$H(i) = -\sum_{j=1}^{K} e_j.log_2(e_j). \tag{2.18}$$

   Misra et al. [84] proposed the use of spectral entropy as a discriminator for automatic speech recognition. Spectral entropy is the estimate of the peakiness of the distribution. Voiced sounds have peaky spectrum and low entropy while noisy and non-speech regions have flatter spectrum and high entropy. Thus voiced and unvoiced speech can be differentiated.

2. **Spectral Centroid**

   The spectral centroid can be defined as a point where the center of mass or center of spectral power is located. It is the measure of the brightness of sound signal. More the centroid, higher the brightness (high frequency) and vice versa. It specifies how the most dominant frequency of the signal changes.

   $$SC_i = \frac{\sum_{k=0}^{K-1} k.|X_i(k)|^2}{\sum_{k=0}^{K-1} |X_i(k)|^2} \tag{2.19}$$

   where SC_i is for i^{th} audio frame.$X_i(k)$ is the amplitude of bin k of DFT spectrum of the i^{th} audio frame and K is frame size. It is used for classification of music genres [5],speech recognition [85], classification of acoustic environment and surveillance of acoustic environment [10].

3. **Spectral Flux**

   The spectral flux determines rate of change of power spectrum of the audio signal. Power spectrum for one frame of the audio signal is compared against the power spectrum from the previous frame. This comparison accounts for Spectral flux. It is a Euclidean distance (2-norm) between the two spectra which are normalized.

$$SF_f = \sum_{k=0}^{K-1} ||X_f(k)| - |X_{f-1}(k)|| \tag{2.20}$$

   where $K$is frame size and index of the frame is f. Audio and speech signals are segmented based on spectral flux [86].

4. **Spectral Roll-off Point (SRP)**

   The spectral Roll-off point is the frequency in power spectrum below which 85% or 95% power is concentrated. It represents the amount of right skewness of power spectrum.

$$SRP = f(N) \quad \text{where} \quad f(N) = N.\Big(\frac{f_s}{K}\Big) \tag{2.21}$$

   where K is frame length and N is 85% or 95%. It is used in the segmentation of music and speech [81], music genre classification [5] and segmentation of audio [87]. For voiced speech, the value of spectral roll-off is low while it is high for unvoiced speech or music.

5. **Mel Frequency Cepstral Coefficients (MFCC)**

   MFCCs are features that are widely used for processing of audio. Earlier Linear Prediction Cepstral Coefficients(LPCC) were used. But these coefficients were replaced by MFCCs because it was discovered that perceived frequency of sound does not change linearly but exponentially. MFCCs are used to model this change that takes place exponentially.

   First pre-processing is performed on the signal which includes framing and windowing. The audio signal is split into frames of about 20ms-40ms so that framed signal is considered pseudo-stationary. Windowing is applied to remove the edge effects. The Discrete Fourier Transform is applied on the frame. Then these Fourier powers are mapped to Mel scale using Mel Filter Banks. This step makes the meaningful frequencies prominent and smoothens the spectrum. Mel Scale gives an indication of what should be the width of Mel filter bank and what should be spacing between two filter banks. The conversion of frequency to mel is given as:

$$mel(f) = 2595 log_10\Big(1 + \frac{f}{700}\Big) \tag{2.22}$$

The resultant of mapping is filterbank energies. The logarithm of filterbank energies is taken because the perceived sound of humans varies logarithmically. Then Discrete Cosine Transform(DCT) is applied to the log filterbank energies to decorrelate the filterbank energies as they overlap. The resultant of DCT is Mel cepstral coefficients but only top 8-13 cepstral coefficients are taken as they contain valuable information about the shape of spectrum and rest coefficients contain redundant information. The equation for DCT of log filter bank energies is given by:

$$c_i = \sum_{k=0}^{K-1} (logS_k).cos(\frac{i\Pi}{K}(k - \frac{1}{2})) \tag{2.23}$$

where c_i is i^{th} cepstral coefficient. K is the number of Mel-filter banks used(20-40). $S_k$is the output of k^{th} filter bank channel. In [88], speaker and speech recognition were performed automatically using MFCC. They are also used in healthcare domain [89], surveillance [90], recognition of music [91] monitoring the traffic parameters like class [61], speed, length of the vehicles etc. and detecting traffic congestion state [20].

2.3.3 Classification Techniques

Classification is the function which assigns the test data to the target classes. So, in other words, classification predicts the outcome for each item in the test data. Classification can be supervised or unsupervised [92]. In Supervised classification, classifier is given the training data containing the items and their respective classes to which each of the items belongs to. This kind of data is called labeled data. The classifier learns through training and finds some distinguishing relationship between item and class. Next, when unseen data called test data is given, the classifier does not know the outcome before but predicts the class based on what it learned during training. The ability to classify the unseen items is called classification accuracy of the classifier. Examples of supervised classifiers are Gaussian Mixture Model(GMM) [93], Hidden Markov Model(HMM) [94], Support Vector Machine(SVM) [16], Artificial Neural Networks(ANN) [63], K-Nearest neighbor(K-NN) [64], Naive Bayes [93].

On the contrary, in unsupervised classification, the classifier is given the data without its outcome called unlabelled data. The classifier clusters the items in data based on similarity among the items. When new unseen data comes, the classifier matches the items with the clusters already formed and put them in one or the other clusters. Various unsupervised classifiers are Self Organising Maps(SOM) [95], Linear Vector Quantization(LVQ) [96], k-means clustering [97].

The classifier must be selected by taking into account computational complexity and classification accuracy. The classifier should take less time to compute result, hence should be computationally simple and consume less power. The classification accuracy must be high so that the classifier is able to classify the items or vectors accurately. Some of the classification algorithms used in earlier works on traffic management are discussed below:

Gaussian Mixture Model (GMM)

The Gaussian Mixture model is categorized as a parametric classifier. In simple words, GMM is the sum of Gaussians. Mathematically, GMM is the weighted sum of means and covariances of different Gaussians. It can be expressed as following equation:

$$p(x) = \sum_{k=1}^{K} w_k g_k(x|\mu_k, \sigma_k) \tag{2.24}$$

where K is number of gaussian components.

The distribution of Gaussians is probability densities that integrates to 1.

$$\sum^{K} w_k = 1, w_k > 1 \tag{2.25}$$

GMM is parameterized by the mean, mixture weights and covariance matrices of all the constituent Gaussians. Gaussian Mixture Models are used for audio classification [98] [99] .

Support Vector Machines(SVM)

SVM is a supervised classifier of machine learning. It is widely used for classification. Given a training set, SVM finds a hyperplane which separates the data set into two classes. The items in training data are represented as points in the plane. Hyperplane, in simple words, if we think of a binary classification, is a line that divides the data set into two classes. Hyperplane is decided by taking into account the distance of data points from the hyperplane. The distance of the data points from the hyperplane should be maximum, while retaining the correct class. The distance between hyperplane and closest points of each class is called margin. More the margin, more the chances of accurate classification. When new test data comes,when the data is projected onto the plane, the side of the hyperplane in which the projections of items are, that is class of those items.

The points which decide the position of hyperplane are called support vectors. In other words, the data points of each class closest to the hyperplane are support vectors.

They are critical elements, as if they are removed or altered, the position of hyperplane will alter. Figure 2.11 describes the hyperplanes, support vectors and margin. Since the hyperplane B has the highest margin from support vectors of both classes , so hyperplane B is the correct one.

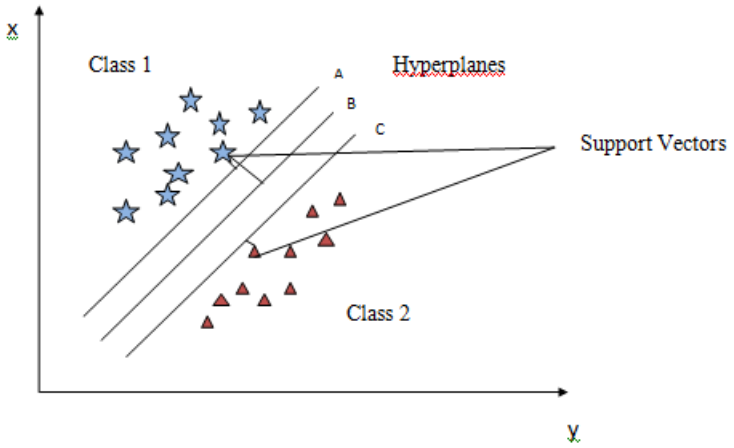

Figure 2.11: Hyperplane in SVM

If the data to be classified is not linearly separable i.e. it is not possible to find a line or plane separating the data points into classes, then the data should be mapped from low dimension to higher dimension using non-linear mapping already decided. The non-linear function used for mapping is also called kernel function. Commonly used and known kernel functions are Gaussian radial basis function, multilayer perceptron, linear, quadratic, polynomial etc. A separating hyperplane is found to classify the higher dimensional data. Non-Linear mapping is depicted in Figure. 2.12

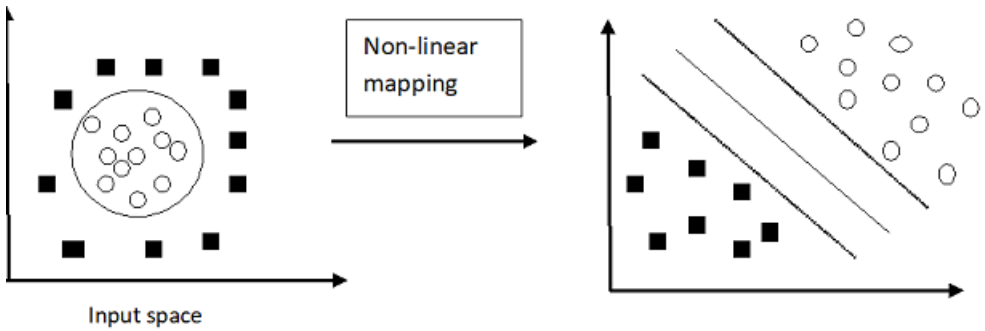

Figure 2.12: Non-Linear Mapping

SVM is a linear classifier basically but it becomes non-linear classifier when non-linear separable data is mapped to higher dimensions using non-linear mapping. SVM was initially developed only for two class problems. But if the problem involves multiple classes, then it is solved using n SVMs. It can be considered a set of binary classification

subproblems. Each SVM out of n SVM divides one class from the other rest of the classes. The decomposition of the problem of multiclass problem into several binary problems is called binarization. There are two techniques employed to solve binarization problem:

1. **One-vs-All(OVA):** In this technique, using SVM one class is separated from all other classes by classifying the examples into two categories. One containing all the examples belonging to that particular class and other containing the examples not belonging to that class. So, there is one SVM for each class.

2. **One-vs-One(OVO):** In this technique, there is one SVM for each pair of classes. Each SVM performs classification between two classes. One class wins in case of each classification task. The final decision about the class of the data item is taken through voting. The score of the winner class increases based on the decision of each classifier. At the end of the classification process,the class which has the maximum score is the class of the data item.

SVMs are widely used for vehicle classification using audio [100] and traffic state congestion detection using audio [20] [31] [16].

k-Nearest Neighbor Classifier (k-NN)

K-NN is applicable to multiclass problems. The data is fed to the algorithm with outcomes and this data with its labels is used to predict the class of new data item. When a new data item is fed, the algorithm find out which of the items of training set are closest to the new unseen data item. There is a specific number chosen of the closest data items and there is a measure of closeness defined. Out of the closest data item, the class to which maximum items belong is considered as target class of the unseen data item. K-NN is employed by researchers for vehicle classification using audio [64].

Artificial Neural Network(ANN)

Artificial neural network is a classification model that works in the same way as biological neural networks. They consist of number of interconnected nodes called neurons. Number of input signals are given to each neuron which produces an output that is forwarded to neuron in next layer as input. Input is received by input layer and output is produced by output layer. Between input layer and output layer, there is one or more layers called hidden layers. Initially, ANN is given random weights of connections. Using input and random weights, output is produced. This obtained output is compared with the desired

output. If they are different then the weights are adjusted using learning techniques which can be supervised, unsupervised or reinforcement learning.

ANNs are widely used for vehicle classification based on audio data [59] [63] [64].

2.4 Convolutional Neural Networks

The convolutional neural network is a deep learning model. It is an artificial neural network that is the variation of feedforward multilayer perceptron networks. It requires very less preprocessing as compared to MLP. It uses mathematical operation called convolution rather than matrix multiplication in at least one of its layers. It preserves the spatial structure in a domain to which it is applied. It was developed primarily for tasks of object recognition. E.g. handwritten digit recognition [101].

2.4.1 Architecture of Convolutional Neural Network

The architecture of CNN has four main types of layers: Convolutional layer, ReLu layer, Pooling layer and Fully Connected layer (Figure 2.13)

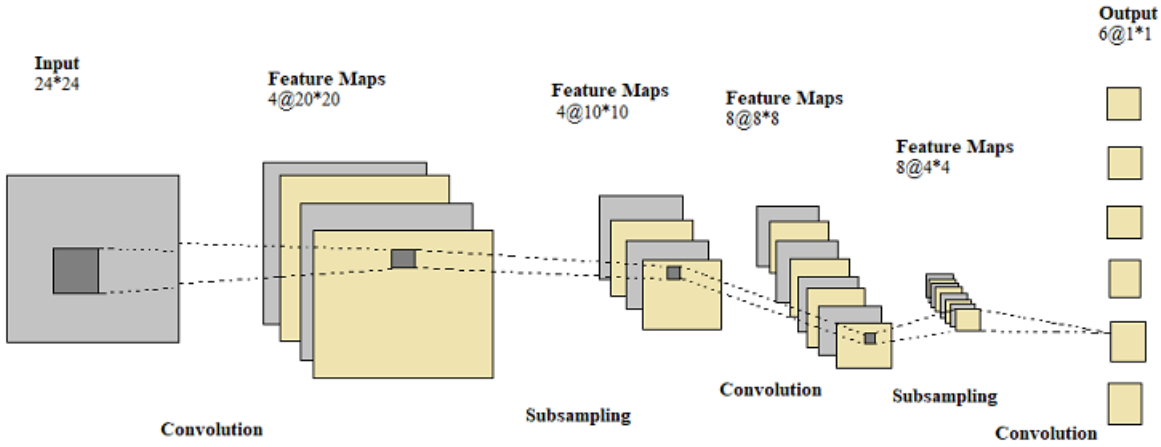

Figure 2.13: Convolutional Neural Network

- ***Convolutional Layer***

  The convolution is a mathematical operation between two functions or signals. An asterisk(*) is used to denote convolution operation. One dimensional version of

convolution is written as:

$$F(t) = (f * g)(t) = \sum_{a=-\infty}^{\infty} f(a)g(t-a) \tag{2.26}$$

Where f is the input and g is the filter.t is the time index and a is the value of time shift. The output of above operation i.e. convolution is activation map of the respective filter used called as feature map.

The convolutional layer is made up of filters and feature maps. Convolutional layer comprises set of learnable filters which are the neurons or kernel of the layer. Filters are the array of numbers called as weights and output a single value within each region. They are used to find out if a particular pattern occurs in input and in which regions. The filter is of smaller width and height than the input and it is convolved with the small region of the input. The region of the input with which filter is convolved is called receptive field. For e.g. In case of image input let input is 32x32x3 then the filter size can be 5x5x3. The filter shifts over the input and is convolved with every position of input. In the end, the array obtained is called feature map, also called activation map. Feature map will be different for each filter. The network will learn filters which are activated when the certain feature is detected. The detected feature can be an edge or color change in case of images and it can be frequency component in case of sound. The output volume in case of the example discussed above is 28x28x3.

The convolutional layer can take input directly if it is the input layer. However if convolution layer is deep inside the network then it will take feature map as input from previous layer.

Important terms related to convolutional layer:

- **Stride:** The amount by which filter shifts is called stride. The value of stride has the effect on the overlapping of receptive fields and output volume. When stride is lets say 1, then filter is moved one pixel at a time over the input volume. This small value of stride results is larger overlapping receptive fields and larger output volumes. As the stride increases, overlapping of receptive field decreases and size of output volumes decreases.
- **Zero Padding:** When filter is applied to the input volume, the output volume is less than the input volume that is spatial dimensions decrease. As more and more convolutional layers are applied the volume will keep on decreasing. But sometimes,it is required to preserve much information about the input so that

required low level features can be extracted. So to get the output of the same dimension as input, zero padding is used. It pads the input with zeros on the border. When the filter is applied to this padded input then output volume is of the same size as input. Size of needed zero padding is to be decided by the following formula

$$Zeropadding = \frac{k-1}{2} \tag{2.27}$$

where k is the size of the filter.

For calculating output size of each convolutional layer, following formula is applied:

$$Outputsize = \frac{f-k+2p}{s} + 1 \tag{2.28}$$

where f is input size, k is filter size, p is the size of zero padding applied and s is stride.

- ***Detector Layer***

  In this layer non-linear activation function such as ReLU (Rectified Linear Units) are applied. This layer makes the output of linear operation(convolution) as non linear. Relu activation function as described above changes all the negative activations to 0. Historically, non-linear functions such as tanh or sigmoid were used but ReLu has much added advantage of increased computational efficiency due to which network is trained much faster. The use of Relu also removes the Vanishing Gradient Problem. It is a problem in which layers train the network very slowly because there is an exponential decrease of gradient through the layers. This layer increases the non-linear properties of the network and also the size of activation units is retained.

- ***Pooling Layer***

  After the convolutional and detector layer, the next layer is pooling layer, also called as downsampling layer. It reduces the spatial size of representation by replacing the output of the network at a certain position with the aggregation operation on the neighborhood. There can be several aggregation operations such as max pooling, average pooling, L2 norm pooling and weighted average pooling. Maxpooling is the most popular which returns the maximum value within neighborhood of values.

The intuitive reasoning behind this layer is that actual location of the feature, if present within the input volume is not as necessary as its relative location with respect to other features. As mentioned above, since the spatial size of representation decreases, so the number of weights to be learned decreases and hence computational cost decreases. It also reduces the problem of overfitting. Overfitting occurs when the model is grown specifically for training examples and it fails to work well for validation and test sets. For e.g. model getting accuracy of 100% in training set but let us say 40-50% in case of test sets is said to be overfitted.

- ***Fully Connected Layer(FC)***

  The fully Connected layer is always the last layer in CNN model. The output of the previous layer (either convolutional or Relu or max-pooling layer) is inputted to this layer and it outputs a vector of length equal to the number of classes into which data is to be classified. The FC layer can use different activation functions-Sigmoid, Softmax etc. The most frequently used activation function is softmax. When it used softmax, each value of the vector is the probability of the class. The class which has the highest value is predicted as the class of the signal.

2.4.2 CNN in Acoustics

CNN has also been employed for classification of sounds. Various hyperparameters are trained to attain high accuracy using CNN. It has been used for classification of environmental sounds [102] [103], classification of bird species using sounds [104]. Various CNN architectures are studied for classification of acoustic scenes [105] and for recognition of audio events [106]. CNN can be combined with other deep learning neural networks like Long short-term memory neural networks for audio classification [107].

2.5 Research Gaps

Since vehicle classification on roads can help to develop several traffic management policies and can help in various applications like automatic toll collection, some technique needs to be provided to perform automatic vehicle classification. Many techniques already exist, but they need some improvement.

- Many intrusive sensing techniques like inductive loop detectors, piezoelectric sensors, magnetic loop detectors have been used for vehicle classification but they have

high installation and maintenance charges and disrupt the traffic while installation [29] [20].

- Many non-intrusive techniques like video cameras, infrared sensors, microwave radars have been used for vehicle classification but they are costly and need to be fixed in one place. Sensors like video cameras are not resistant to occlusion and harsh weather conditions [31] [28].
- Acoustic sensors used till now for vehicle classification are installed along the road at a fixed place [61] [65].
- Researchers have used only human-defined features from acoustics for vehicle classification [19] [61].
- There is no use of crowdsourcing in any of the techniques proposed for vehicle classification using acoustics.
- There is no probing in the field of transfer learning in case of vehicle classification. Whether the model pretrained on dataset of one domain be used in another domain is not explored.

We have proposed a new approach of inferring vehicle category through acoustic recordings of vehicles. Smartphone based acoustic sensors are employed to capture the sounds of vehicles. We have investigated the efficacy of Convolutional Neural Network(CNN), SupportVector Machines(SVM) and combination of CNN and SVM for vehicle classification. The three models are tested on the self collected dataset [108] of 4789 audios of vehicles. For combination of SVM and CNN, features are extracted using deep learning model i.e. CNN and classification is performed using SVM. For the other two models, MFCCs features are extracted and fed to the classifiers. The proposed approach surpasses the human-defined features based approaches for vehicle classification.

Chapter 3

Methodology

The research gaps described in previous chapter highlight the major challenges in domain of vehicle classification. A new system has been proposed to overcome these challenges and fulfill the objectives listed in the section 1.5. This chapter describes the system design of new proposed approach. The first section of this chapter gives a detailed description of the proposed method. The second section gives the detailed description of the approaches compared with the proposed approach and experiments performed to determine the efficacy of the deep learning model for feature extraction are listed in next section. We have also explored if fusing human-defined features with Convolutional Neural Network (CNN) based features can give better results. The experiments related to feature fusion are illustrated in the last section.

3.1 System Overview

Figure 3.1 describes the whole process of the proposed vehicle classification system using sounds of vehicles acquired through smartphones. Acoustic sensors present in smartphones are triggered based on accelerometer which detects orientation change. First, the sounds acquired from the vehicles are preprocessed, then features are extracted from each audio file using CNN. The features are signatures of vehicle types. Then these signatures are given as input to the classifier i.e. Support Vector Machine (SVM) which classifies the type of vehicle. This information is communicated to the server. The server may be on the side of traffic management policymakers or it can be on the side of applications management staff like automatic toll collection, parking system designs etc. Following subsections explain the process in detail.

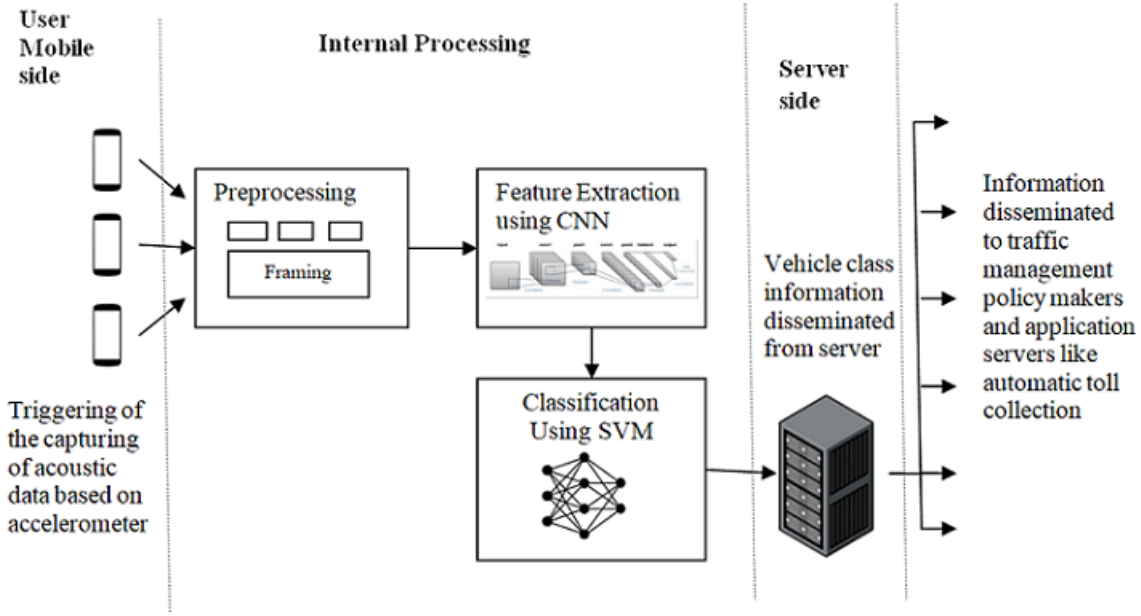

Figure 3.1: System Overview of Proposed approach

3.1.1 Preprocessing

Preprocessing of audio signals is required to remove the noise and irrelevant chunks of the audio signal. It includes framing i.e. dividing the audio signal into series of frames and application of Discrete Cosine Transform (DCT) on each frame of audio signals.

Framing

Framing is splitting the acoustic signal into the succession of frames. As discussed in section 2.3.1, audio signal is continuously varying signal. So it is difficult to extract features from moving signal. Hence, the signal is bifurcated into frames so that each frame is considered quasi-stationary. For different applications, different frame sizes are considered the best. For traffic related applications such as traffic congestion detection, accident detection and vehicle classification, higher frame size is needed as compared to the speech-related applications [20].

Further, we have taken overlapping of frames into account to keep track of events and features which occur at the discontinuity. Overlapping size is the amount by which one frame coincides partly with another frame.

3.1.2 Feature Extraction

Feature Extraction is the process of extracting the relevant characteristics or information that describes the audio signal. Features act as the signature of vehicles. Various types of features can be extracted from the audio signal. These can be temporal fea-

tures or spatial features. As discussed in section 2.3.2, some of the temporal features are Zero Crossing Rate(ZCR), Short Time Energy(STE), Autocorrelation, Root Mean Square (RMS) and Energy Entropy(EE). Some of the spectral features are Spectral Roll-Off, Spectral Entropy, Spectral Flux, Spectral Centroid and Mel Frequency Cepstral Coefficients(MFCCs). These are extracted from each audio recording and fed to the classifier which classifies the audio signals into various classes of vehicles.

Feature Extraction can be either human-defined or through some features extracting model. We have employed CNN to extract features in the proposed off the shelf CNN features based approach for vehicle classification using acoustics. This proposed approach is compared with the human-defined features based machine learning and deep learning approaches. In human-defined features based deep learning approach, MFFCs are given as input to the CNN which classifies the sound recordings into one of the categories of vehicles. Similarly, in human-defined features based machine learning approach, MFCCs are given as input to the SVM.

Mel Frequency Cepstral Coefficients (MFCC)

MFCCs are features that are widely used for processing of audio. Section 2.3.2 describes MFCCs in detail. We have taken into account top 13 coefficients for our experiments as top 13 coefficients contain the non-redundant information in case of traffic related applications [31].

Feature Extraction using CNN

CNN, as discussed in section 2.4.1, comprises of the number of layers- Convolutional layer, Relu layer, Pooling layer, Fully Connected layer. The last layer i.e. fully connected layer predicts the probabilities for an object to belong to a specific class. Before this, the layers in CNN extract features from the audio signal, features that are specific to each class of vehicles. We have extracted features from the second last layer of CNN. Then these features are fed as input to the SVM classifier.

3.1.3 Classification

The feature vectors extracted are fed to the classifier. The classifier classifies the audio signal as belonging to the specific class by working on its feature vectors. We have experimented with two classifiers- CNN (section 2.4) and SVM (section 2.3.3). The audio frames are split into three sets- training set, validation set and testing set. When training set is fed to the model, then feature vectors along with their respective labels are

given as input to the classifying model. The model is trained to classify the audio signals to different classes of vehicles. When the model is trained, then the model is validated using validation set. The test data is given as input without the labels and model is tested for classification accuracy. The accuracy of the model is determined by the fraction of correctly predicted labels.

3.2 Approaches Compared

We have compared the proposed approach (CNN and SVM) with the two approaches-human-defined features based machine learning approach (MFCCs and SVM), human-defined features based deep learning approach (MFCCs and CNN) (Figure 3.2). These approaches are explained in the following subsections.

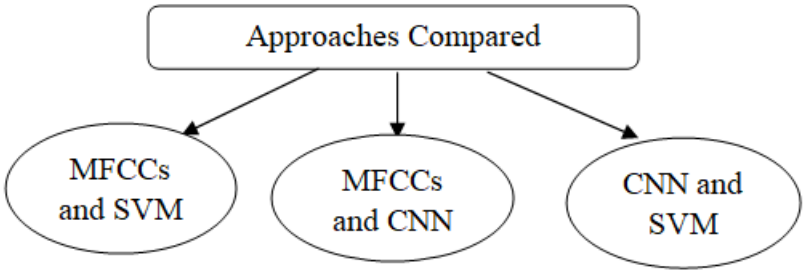

Figure 3.2: Approaches Compared

3.2.1 Human-defined features based machine learning approach (MFCCs and SVM)

In human-defined features based machine learning approaches, human-defined features from acoustics are given as input to the machine learning classifier. This approach is commonly used for various applications. Either spectral or temporal features or both are extracted from acoustic recordings and are given as input to the classifier i.e. Support Vector Machine [16], Artificial Neural Network [63], K-Nearest Neighbour Classifier [64], Multilayer Neural Network [65] etc. We have used MFCCs for this approach. 13 MFCCs coefficients are given as input to the SVM. As described in section 2.3.2, top 8-13 MFCCs coefficients are enough and contains non-redundant information. In case of traffic related applications, 13 MFCCs coefficients are sufficient [31]. The steps followed for this approach are given below:

- Each audio signal undergoes framing and 13 MFCCs are extracted from each frame.
- The MFCCs from all the frames of an audio signal are combined through average to get a single feature vector corresponding to each audio file.
- The labeled feature vectors extracted from training data of audio files are used to train the SVM. The trained model is then used to classify the new audio files.

This approach is illustrated in Figure 3.3.

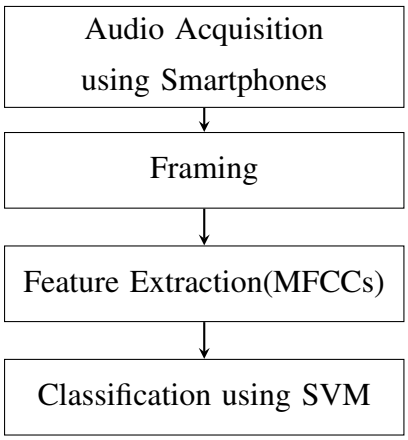

Figure 3.3: Human-defined features based machine learning approach (MFCCs and SVM)

3.2.2 Human-defined features based deep learning approach (MFCCs and CNN)

In human-defined features based deep learning approach, human-defined features from audio recordings are given as input to the deep learning model [102]. MFCCs extracted from the audio recordings are given as input to the CNN (section 2.4) [103]. Similar to the above approach, we have extracted only 13 MFCCs features which are given as input to the CNN. The labeled feature vectors extracted from training data of audio files are used to train the CNN. Then, CNN classify the new audio recordings of the vehicles. The steps followed for this approach are given below:

- Each audio signal undergoes framing and 13 MFCCs are extracted from each frame.
- The MFCCs from all the frames of an audio signal are combined through average to get a single feature vector corresponding to each audio file.

- These feature vectors are used to train CNN and CNN classifies the new audio files into different vehicle categories.

This approach is illustrated in Figure 3.4.

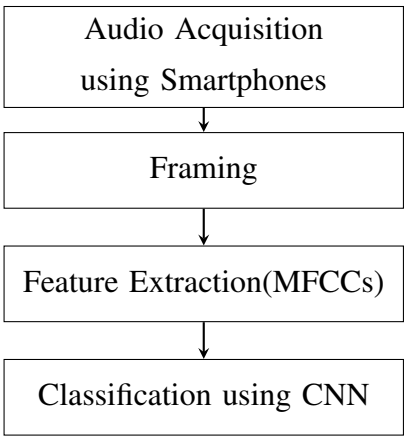

Figure 3.4: Human-defined features based deep learning approach (MFCCs and CNN)

3.2.3 Proposed off the shelf CNN features based approach (CNN and SVM)

In the above discussed approaches, human-defined features are given as input to the classifier, either machine learning classifier or deep learning classifier. We have proposed to combine deep learning model and machine learning model. Since, the deep learning model can be used end-to-end as it can extract features and classify the acoustic recordings itself, so we have proposed to use deep learning model i.e. CNN for extracting features. These features are fed to the machine learning classifier i.e. SVM.

- In this approach, each audio frame undergoes discrete cosine transformation(DCT). This is required to tranform the audio recording into frequency domain. Then the DCTs of all audio frames is combined to get a single vector for each audio file.

- The DCT vectors corresponding to each audio file are given as input to CNN which extracts features for each audio file.

- The feature vectors extracted using CNN are given as input to SVM for training. SVM categorizes the unseen audio files into one of the vehicle classes.

CNN is used as feature extractor and SVM is used for vehicle classification. The overview of this approach is illustrated in Figure 3.5.

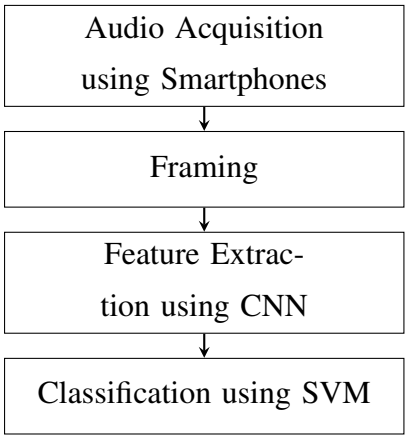

Figure 3.5: Proposed off the shelf CNN features based approach (CNN and SVM)

3.3 Determination of efficacy of CNN for feature extraction

Convolutional Neural Network, as described in section 2.4.1 consists of layers- Convolutional layer, ReLu layer, Pooling layer and Fully Connected layer. The number of layers play an important role in performance of the model. Experiments are performed with the two-layered CNN and four-layered CNN to determine if number of layers affect the performance of the proposed approach. Then, efficacy of CNN is determined by varying different parameters like filter size, stride length, number of filters, pool size and kernel functions .

3.3.1 Model 1: 2-layered CNN Model and classification using SVM

The input reshaped into 2*2 matrix is convolved with filters. The dimensions of output of convolutional layer are reduced by one max pooling layer. In between, activation function Relu is used. Features are extracted from the max pooling layer of CNN. These features are used to train SVM model. The number of features extracted from CNN varies with the number of filters and stride length. The number of features corresponding to number of filters and stride length is discussed in Chapter 4. One of the experimental setup is shown in Figure 3.6.

Experiments performed on Model 1

Following experiments are performed to optimize the model by tuning the different hyperparameters.

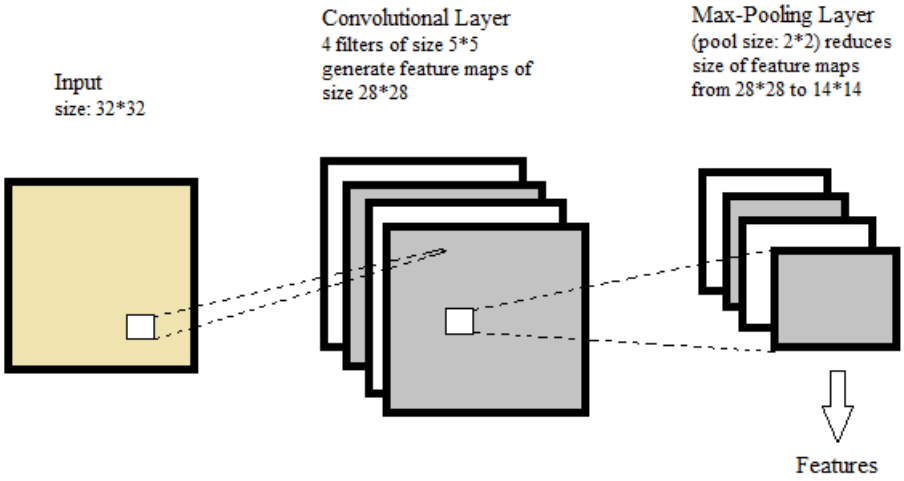

Figure 3.6: Feature Extraction using CNN

1. **Number of filters**

   Changing the number of filters can affect the performance of the vehicle classification system. So, number of filters are varied from 4 to 256 including 4, 8, 16, 32, 64, 128 and 256 filters.

2. **Stride length**

   Stride controls the amount by which filter shifts across the input. Stride is chosen so that the output of the network is integer rather than fraction. If the stride length is small then the receptive fields produced after filter convolution are more overlapping as compared to the case when the stride length is large [109]. As discussed in section 2.3.1, there can be number of events which occur at discontinuity while framing. So, overlapping needs to be considered. So, different stride lengths are considered-1*1, 2*2 and 4*4.

3. **Different kernels of SVM**

   There are various types of kernels of SVM like Linear, Radial Basis Function, polynomial, sigmoid function etc. Different kernel functions have different effects on the accuracy of SVM. We have experimented with three types of kernels of SVM: Linear, Radial Basis Function (RBF) and Sigmoid function.

4. **Pool size in max-pooling layer**

   Pool size, as explained in section 2.4.1, can be varied according to the need for applications. Pooling layer is used to reduce dimensions of the output of convolutional layer and can reduce the number of parameters to be learned by the model.

We have performed experiments with different pool sizes i.e. 2*2, 3*3, 5*5, 7*7, 9*9 and 11*11.

5. **Effect of Size of Input**

   As, different input sizes can have varying effect on the performance of the CNN so variations of input sizes are tried- 28*28, 32*32, 36*36, 40*40, 44*44, 48*48, 52*52.

6. **Size of filter in convolutional layer**

   Size of filter is an important parameter that can affect the accuracy of the classification system. So, the size of filters are varied (3*3, 5*5, 7*7 and 9*9 and 11*11).

3.3.2 Model 2: 4-layered CNN Model and SVM

The input of size 48*48 is convolved with filters of size 9*9. The dimensions of the output of convolutional layer is reduced by one max pooling layer. In between, activation function Relu is used. Then, one more convolutional layer and max-pooling layer are used. Features are extracted from the max pooling layer of CNN. These features are used to train SVM model. In this model, the number of features extracted from CNN varies with the number of filters only as described in following chapter. Figure 3.7 illustrates the basic architecture of four-layered CNN model used for experiments.

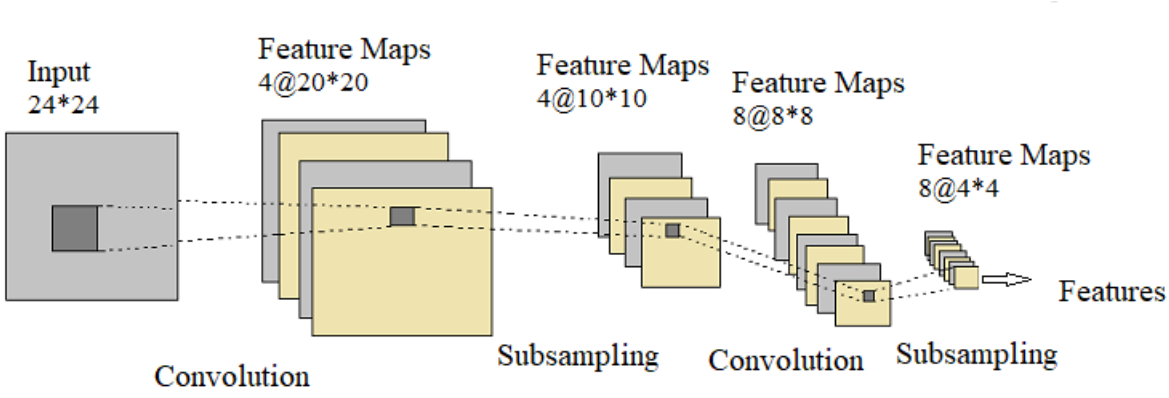

Figure 3.7: 4-layered CNN model

Two different kinds of experimental setups are used in case of Model 2.

1. **Experimental Setup 1**

   Here we have the following parameter values:

   - First Convolutional layer has filters with dimension of 9*9 (conv1).

- Stride length used is 1*1.
- First max-pooling layer has pool size of 7*7 (maxpool1).
- Second Convolutional layer has filters with dimension 3*3 (conv2).
- Second max-pooling layer has pool size of 3*3 (maxpool2).

2. **Experimental Setup 2**

   Here we have the following parameter values:

   - First Convolutional layer has filters with dimension of 9*9 (conv1).
   - Stride length used is 2*2.
   - First max-pooling layer has pool size of 5*5 (maxpool1).
   - Second Convolutional layer has filters with dimension 3*3 (conv2).
   - Second max-pooling layer has pool size of 2*2 (maxpool2).

In both the setups of model 2, various hyperparameters are varied to find their effect on the performance of the CNN. Table 3.1 illustrates different hyperparameters and their corresponding values which are used for experiments.

Table 3.1: Hyperparameters for Model 2

Hyperparameters	**Values**
Stride Length	1*1, 2*2
Dimensions of filter	3*3, 5*5, 7*7, 9*9, 11*11
Different kernel functions	Sigmoid, Linear and Radial Basis Function
Number of filters	4, 8, 16, 32, 64, 128

3.4 Feature Fusion

We have explored feature fusion to find out if the fusion of human-defined features and features extracted using deep learning model can improve the performance of the vehicle classification system. The features from CNN and MFCCs of each audio file are combined. Number of features from intermediate layer of CNN vary with the change in number of filters and stride length. The model 1 of the proposed approach described in section 3.3.1 is used for feature fusion experiments. Figure 3.8 illustrates the process of feature fusion.

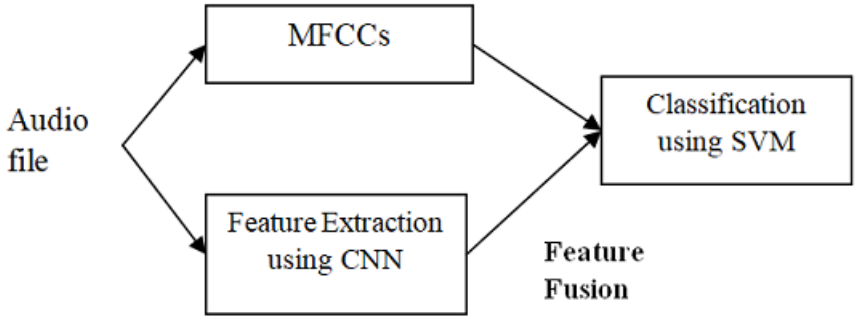

Figure 3.8: Feature Fusion

Hyperparameters of CNN are varied in case of feature fusion. Table 3.2 illustrates the hyperparameters and their corresponding values that are used for experiments.

Table 3.2: Hyperparameters for Feature Fusion

Hyperparameters	**Values**
Stride Length	1*1, 2*2, 4*4
Number of filters	4, 8, 16, 32, 64, 128, 256
Pool size in max-pooling layer	2*2, 3*3, 5*5, 7*7, 9*9, 11*11
Dimensions of filter	3*3, 5*5, 7*7, 9*9, 11*11
Different kernel functions	Sigmoid, Linear and Radial Basis Function

3.5 Summary

In this chapter, we have explained the approaches which are compared for vehicle classification using acoustics. The process followed for human-defined features based machine learning approach (MFCCs and SVM), human-defined features based deep learning approach (MFCCs and CNN) and the proposed approach (CNN and SVM) are described in detail. It also gives a detailed overview of the set of experiments that are performed to attain better performance of the proposed approach. Then, the methodology followed for feature fusion i.e. fusing the human-defined features and off the shelf CNN based features is also discussed. Certain experiments perfomed to tune the hyperparameters of the proposed model for feature fusion are explained in detail in this chapter.

Chapter 4

Results And Discussions

4.1 Introduction

We have proposed a new method for vehicle classification using sounds collected using smartphones based acoustic sensors. Acoustics have been employed in several applications. As of now, acoustics have been extensively used in speech, music, environmental scene recognition etc. (section 1.1.1). In the field of traffic, acoustics have been used for traffic congestion detection. But the little work has been done in classification of vehicles on road. Environmental scene recognition using acoustics is dealt with in DCASE Challenge 2013 [110] and 2016 [111].

We have modelled a combination of Convolutional Neural Network (CNN) and Support Vector Machines (SVM) for vehicle classification. An off the shelf CNN based feature extraction approach has been employed and classification is performed using SVM. In this chapter, we have made a comparative analysis of performance of human-defined features based machine learning approach (MFCCs and SVM), human-defined features based deep learning approach (MFCCs and CNN) and off the shelf CNN features based approach (CNN and SVM) for vehicle classification. The details of dataset and the set up done for experiments are discussed in first section of this chapter. In the subsequent sections, detailed analysis of various approaches is presented. Then experiments performed for determining the adequate filter size, pool size, number of filters, kernel function and stride length that give the best accuracy are described.

4.2 Capturing Sounds of Vehicles

The main motive of the research is to develop an efficient model that will classify vehicles on roads. The sounds captured from vehicles using smartphones are used to train the model which will help in categorizing the vehicles.

4.2.1 Types Of Vehicles

There are several types of public transport vehicles. The classes of vehicles can be car, truck, train, bus, aeroplane, three-wheeler, bike, military vehicles etc. On the other hand, the vehicles can also be categorized as heavy vehicles, intermediate vehicles and light weight vehicles [112]. In this work, we have classified sounds into five public transport vehicle categories-car, bus, train, aeroplane and three-wheeler (Figure 4.1).

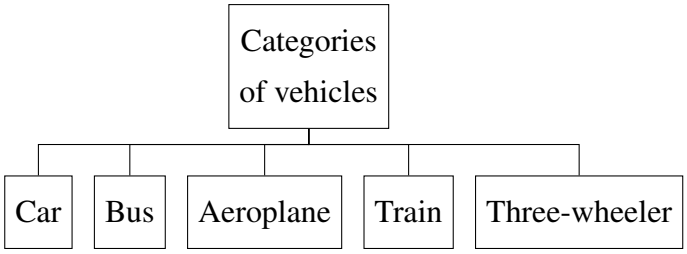

Figure 4.1: Classes of Vehicles

4.2.2 Dataset

We have used the self collected dataset [108] for vehicle classification. The dataset contains 4789 recordings of five vehicle classes-car, bus, aeroplane, train and three wheeler. The number of recordings for each vehicle category is shown in Table 4.1.

Table 4.1: Number of recordings for each category of vehicle in dataset

Vehicle Category	**No. of recordings**
Car	850
Bus	1104
Aeroplane	542
Train	1223
Three-wheeler	1070

4.2.3 Dataset Characterstics

The dataset used for experiments has following characterstics:

- The duration of each recording is 30 seconds approximately.
- Each recording belong to a particular class of vehicle-car, bus, train, aeroplane or three-wheeler.
- Ground truth is manually labelled.
- Sampling rate is chosen to be 16kHz for recordings. Higher sampling rate increases the cost of computation.
- Uncompressed PCM (.wav) format is loseless. Hence, all the recordings are recorded in .wav format.

4.2.4 Experimental Setup

The class labels for each of the vehicle categories used in experiments are shown in Table 4.2.

Table 4.2: Vehicle categories and respective labels used for experiments

Vehicle Category	Class Label
Car	0
Bus	1
Aeroplane	2
Train	3
Three-wheeler	4

For experiments, 80% of the data is used for training, 10% of the data for validation and 10% for testing. All the three sets-training, validation and testing sets are chosen randomly.

4.3 Performance Metrics

There are number of performance metrics in order to evaluate how good a model is: Precision, Accuracy, Recall, F1 score etc. In this work, metric accuracy is used for evaluation of model. Confusion matrix is discussed which forms the base of the performance metrics. Confusion matrix is a table that gives a vivid description of performance of a model

used for classification. It consists of four terms- True positive, False positive, True negative and False negative. Confusion matrix is illustrated in Figure 4.2. The description of the terms used in confusion matrix is given below:

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Figure 4.2: Confusion Matrix

4.3.1 True Positive(TP)

True positive is the number of positive values which are predicted correctly. In other words, values of both actual class and predicted class are yes.

4.3.2 False Positive(FP)

False positive is the number of negative values which are predicted positive. In other words, value of actual class is no and value of predicted class is yes.

4.3.3 True Negative(TN)

True Negative is the number of negative values which are predicted correctly. In other words, values of both actual class and predicted class are no.

4.3.4 False Negative(FN)

False negative is the number of positive values which are predicted negative. In other words, value of actual class is yes and value of predicted class is no.

4.3.5 Accuracy

Accuracy is the proportion of the number of observations predicted correctly to the total number of observations.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

4.3.6 Precision

Precision is the proportion of number of correctly predicted positive observations to the total predicted positive observations. It is also called Specificity.

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

4.3.7 Recall

Recall is the proportion of number of correctly predicted observations to the total number of positive observations in class. It is also called Sensitivity.

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

4.3.8 F1 score

F1 Score is the harmonic mean of Recall and Precision. Since, it takes both false positive and false negative into consideration, so F1 score is more meaningful as compared to accuracy. It is useful in case of class distribution that is uneven.

$$F1Score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \tag{4.4}$$

4.4 Comparative Analysis of three approaches

As discussed in section 3.2, the three aproaches are compared for vehicle classification. This section explains the results of all the approaches and different experiments listed in section 3.3. Each audio file undergoes framing with the frame length of 8192 and hop length of 2048. Experiments are performed with training set size of 0.8, validation set size of 0.1, test set size of 0.1. The hyperparameters C and gamma are tuned to get the validation accuracy same as training set accuracy. We tried with the different values of C (1, 10, 100, 1000) and gamma (0.1,0.01,0.001). C= 100 and gamma= 0.01 are the finally tuned parameters. Values of C beyond 100 and gamma beyond 0.01 do not give favourable results. The best results for each approach are listed in following subsections.

4.4.1 Human-defined features based machine learning approach (MFCCs and SVM)

For each audio frame, 13 MFCCs features are extracted. These features are used to train the machine learning classifier i.e. Support Vector Machine (section 3.2.1). Experiments are performed with different kernel functions- Radial Basis function, Linear and Sigmoid. Radial Basis function gives the best accuracy. The classification accuracies corresponding to different kernel functions of SVM is shown in Table 4.3.

Table 4.3: Classification Accuracy in approach MFCCs and SVM

kernel function	**Classification Accuracy**
Linear	62.00
Rbf	72.65
Sigmoid	54.07

4.4.2 Human-defined features based deep learning approach (MFCCs and CNN)

From each audio frame, 13 MFCCs features are extracted and are given as input to CNN (section 3.2.2). CNN is trained using categorical cross entropy as the loss function. The size of mini-batch in each epoch is 32. The weights are updated using adam after the mini-batch is processed. The first layer contains 256 neurons and input size is 13. Then there is another layer of 256 neurons and finally a fully connected layer. As an activation function, Relu is applied in between. Number of epoches are 20. There are 5 softmax nodes in the last layer corresponding to the number of classes. The CNN gives an accuracy of 79.97% on the test set.

4.4.3 Proposed off the shelf CNN features based approach (CNN and SVM)

CNN and SVM are combined in proposed approach. Raw audio signals undergo framing and Discrete Cosine Transformation (DCT). The resultant of DCT applied over each audio signal are given as input to CNN. CNN extracts features from audio signals and then SVM classifies the audio signal into one of the five vehicle categories (section 3.2.3). First, two-layered CNN model is employed and various hyperparameters like filter size, stride length, kernel function, number of filters and pool size are tested. Then, experiments are performed with four-layered CNN models and the same parameters are tested for the four-layered CNN model.

Model 1: 2-Layered CNN Model and classification using SVM

The details of model 1 are described in section 3.3.1. The number of features corresponding to number of filters and stride length for model 1 are depicted in Table 4.4.

All the findings related to proposed model using 2 layers are listed below:

Table 4.4: Number of features corresponding to number of filters and stride length (Model 1)

	Stride length		
No. of filters	1*1	2*2	4*4
4	64	16	4
8	128	32	8
16	256	64	16
32	512	128	32
64	1024	256	64
128	2048	512	128
256	4096	1024	256

1. **Effect of Number of filters**

   Experiments are performed by increasing the number of filters from 4 to 256 (4, 8, 16, 32, 64, 128 and 256). The classification accuracy increases uptil 8 filters. Then the accuracy becomes almost constant with minor change. This is because, increasing number of filters increase the representational capacity of the model [101]. But, at some point , increase in number of filters do not make a change in accuracy. So, two-layered CNN used for vehicle classification gives the best accuracy at 8 filters. Further, increase in number of filters just increase the time and memory cost with no significant improvement in performance. Figure 4.3 illustrates the effect of the number of filters on the performance of the model.

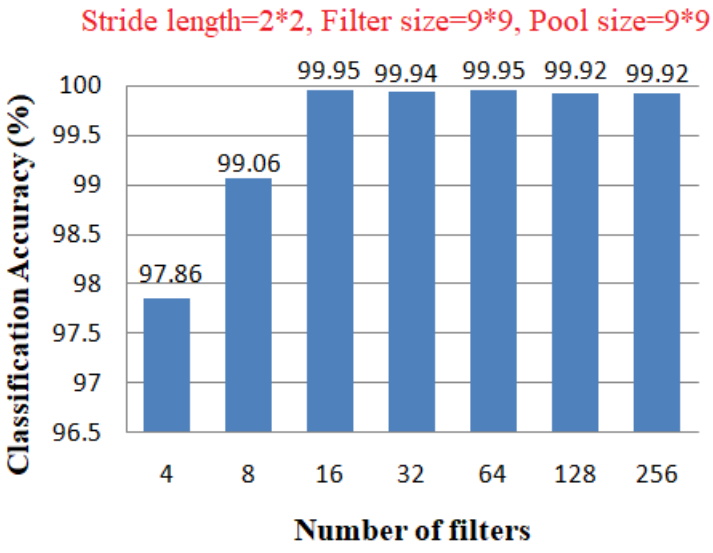

Figure 4.3: Effect of number of filters on the performance of model (Model 1)

2. **Effect of Stride length**

The stride length of 2*2 gives good results. So, overlapping through stride of 2*2 is sufficient to produce effective results. In case of 4 and 16 filters, increase in stride length from 2*2 to 4*4 leads to poor performance but in case of 8 and 32 filters, the change in stride from 2*2 to 4*4 does not have much effect on accuracy of the model. So, choosing the stride length depends entirely on the type of dataset. Figure 4.4 and Table 4.5 illustrate the effect of stride length on the classification accuracy of the model.

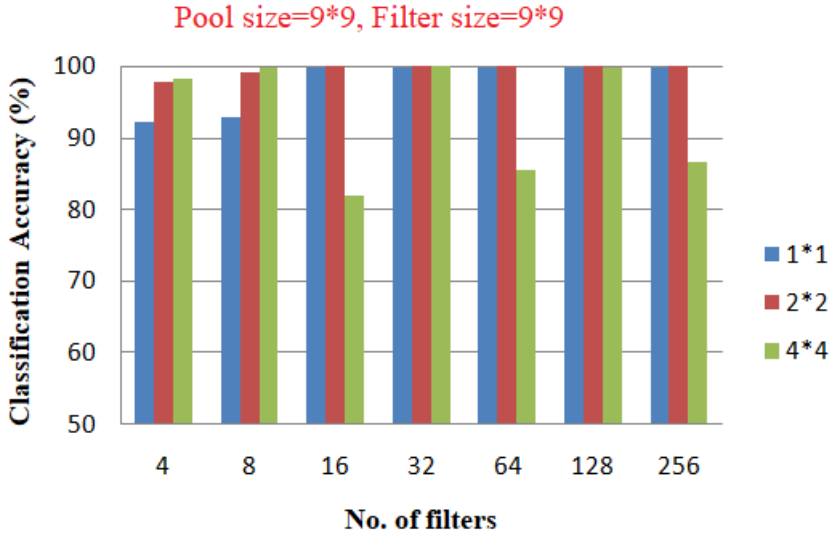

Figure 4.4: Effect of stride length (Model 1)

Table 4.5: Effect of Stride length on classification accuracy (Model 1)

	Stride length		
No. of filters	1*1	2*2	4*4
4	92.22	97.86	78.13
8	92.85	99.06	99.79
16	99.74	99.95	81.94
32	99.90	99.94	99.95
64	99.89	99.95	85.56
128	99.90	99.92	99.85
256	99.87	99.92	86.66

3. **Analysis of different kernels of SVM**

For 4 filters, RBF gives the best classification accuracy and for 8, 16, 32, 64, 128 and 256 filters, linear and sigmoid functions give better classification accuracy as compared to RBF. The results of various kernel functions are listed in Table 4.6.

Figure 4.5 describes the variation in classification accuracy of the model with the change in kernel function of SVM.

Table 4.6: Analysis of different kernels of SVM (Model 1)

	Kernel function		
No. of filters	Linear	RBF	Sigmoid
4	74.48	97.86	74.48
8	99.06	94.89	99.06
16	99.95	97.18	99.95
32	99.94	80.74	99.94
64	99.9	81.16	99.9
128	99.93	80.66	99.93
256	99.94	81.1	99.94

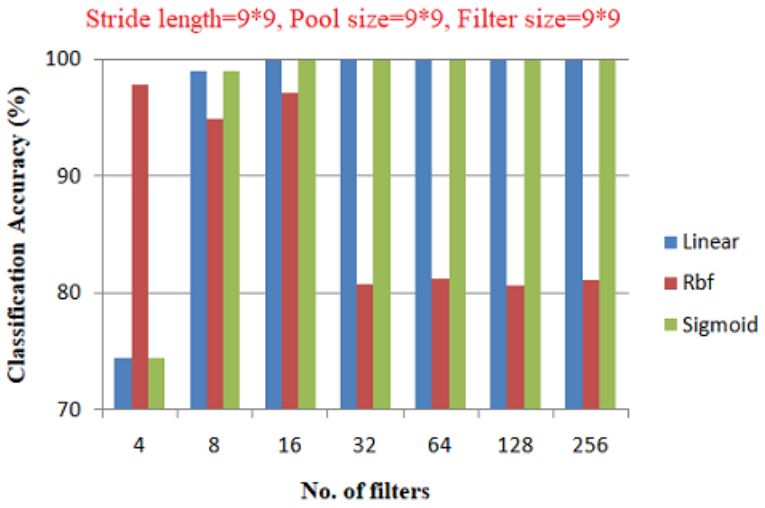

Figure 4.5: Analysis of different kernels of SVM (Model 1)

4. **Effect of pool size in max-pooling layer**

The pool size of 9*9 yielded the best accuracy. The pooling layer reduces the number of parameters and computations and prevent overfitting [109]. But making the pool size much larger can reduce the effectiveness of model as important parameters required for classification can be lost. In this model, 9*9 pool size gave the best results in terms of controlling overfitting, performance without losing the ability of the model to represent itself. Figure 4.6 describes the effect of pool size in max-pooling layer on classification accuracy of the model.

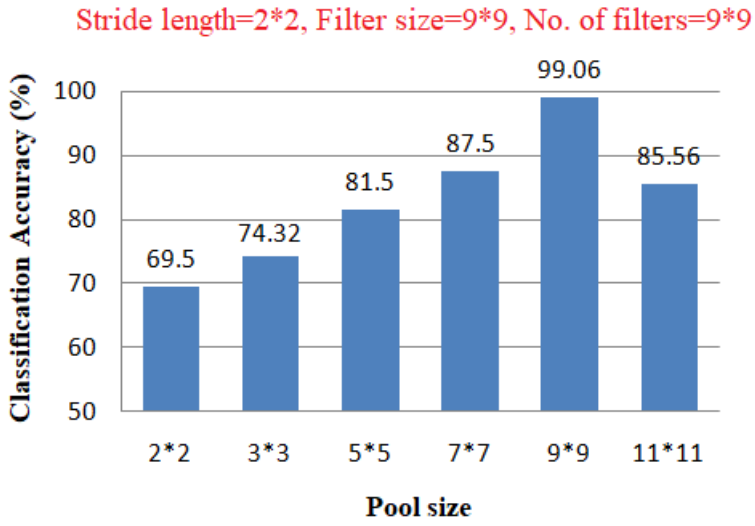

Figure 4.6: Effect of pool size in max-pooling layer on classification accuracy (Model 1)

5. **Effect of Size of Input**

The model performs well for input size of 48*48. Accuracy increases from 28*28 to 48*48 but then the accuracy starts decreasing. In nutshell, by increasing the input size, performance of the model follows the positive gradient [113] till the saturation point, which in this case is 48*48. Then the accuracy starts decreasing with further increase in input size. Figure 4.7 illustrates the effect of the size of input on performance of model.

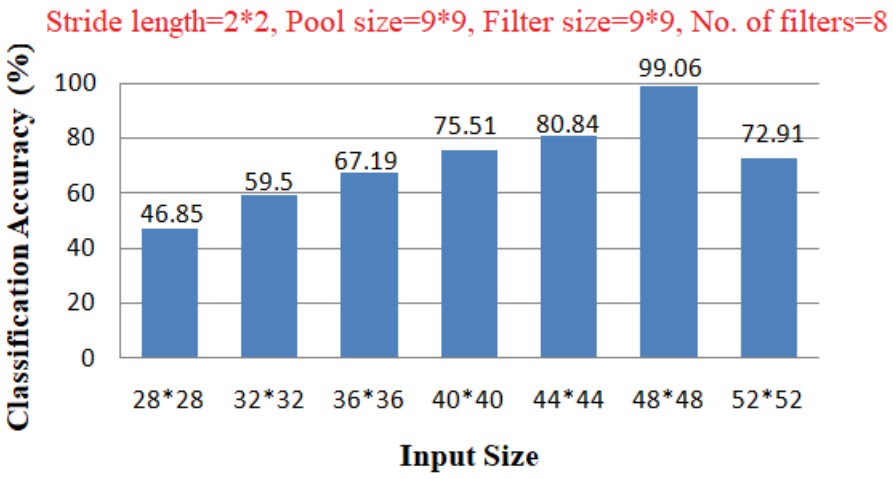

Figure 4.7: Effect of size of input on classification accuracy (Model 1)

6. **Effect of Size of filter in convolutional layer**

The filter size of 9*9 gives the best results. This is because, increase in width of kernels (filters) increases the number of parameters to be learned [101]. Since, the

number of parameters increase, so the model can perform more accurately with the least chances of error. But, increasing the filter size further led to increased run-time because as the number of parameters increases, the requirement of memory for parameter storage also increases. So, 9*9 yields the best results in terms of performance with acceptable cost of memory and run-time. Figure 4.8 illustrates the effect of the size of filters in convolutional layer on performance of the model.

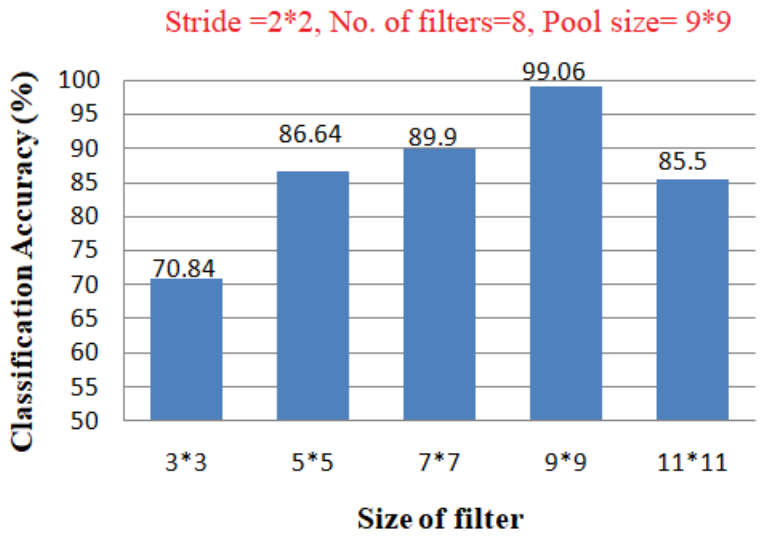

Figure 4.8: Effect of size of filters on classification accuracy (Model 1)

Best Combination of parameters for Model 1: 2-layered CNN model and classification through SVM

Table 4.7 illustrates the best combination of parameters for proposed approach in which two-layered CNN is used and SVM is used for classification. The model 1 of proposed approach gives an accuracy of 99.06% with the best combination of hyperparameters.

Table 4.7: Best combination of hyperparameters for Model 1

Hyperparameters	**Values**
Filter size	9*9
No. of filters	8
Pool size	9*9
Stride Length	2*2
Kernel function	Linear or Sigmoid

Model 2: 4-layered CNN Model and SVM

Model 2 and its experimental setups are described in section 3.3.2. The number of features extracted from CNN varies with the number of filters only as depicted in Table 4.8. The results of experiments performed to determine the best combination of parameters for model 2 are listed below.

Table 4.8: Number of features corresponding to number of filters and stride length (Model 2)

No. of filters	**Stride length (1*1 or 2*2)**
4	4
8	8
16	16
32	32
64	64
128	128
256	256

1. **Effect of Change in Stride length**

   Change in stride length has no considerable effect on accuracy of the model. Figure 4.9 illustrates the effect of change in stride length on the classification accuracy of the model. So the stride length of 1*1 is good for this four-layered CNN model and classification through SVM (Table 4.9).

Figure 4.9: Effect of change in Stride length on performance (Model 2)

Table 4.9: Effect of stride length on classification accuracy (Model 2)

No. of filters	**Stride length**	
	1*1	2*2
4	87.68	89.30
8	99.29	99.37
16	99.63	99.46
32	99.74	99.69
64	99.74	99.90
128	99.84	99.89
256	99.84	99.88
512	99.90	99.89

2. **Effect of change in dimensions of filter**

   The filter size of 9*9 gives the best results. Figure 4.10 illustrates the effect of size of filters on performance of the model.

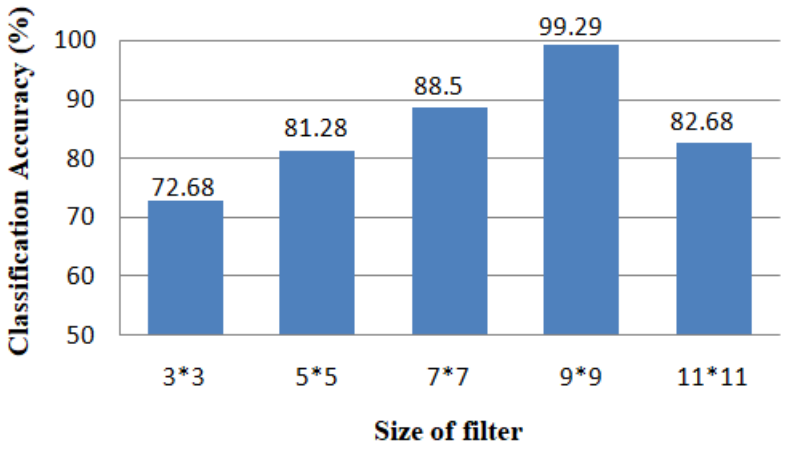

Figure 4.10: Effect of size of filters on classification accuracy (Model 2 setup 1)

3. **Analysis of different kernel functions of SVM**

   For 4 filters, RBF gives good results but for all other number of filters, linear and sigmoid kernel functions give the best results. Figure 4.11 and Table 4.10 illustrate how the performance of the model varies with different kernels of SVM.

Table 4.10: Analysis of kernel functions of SVM (Model 2 Setup1)

No. of filters	**Kernel function**		
	Linear	Rbf	Sigmoid
4	53.44	87.68	53.44
8	99.29	95.82	99.29
16	99.63	78.34	99.63
32	99.74	79.8	99.74
64	99.74	79.12	99.74
128	99.84	79.59	99.84
256	99.84	79.96	99.84
512	99.9	79.85	99.9

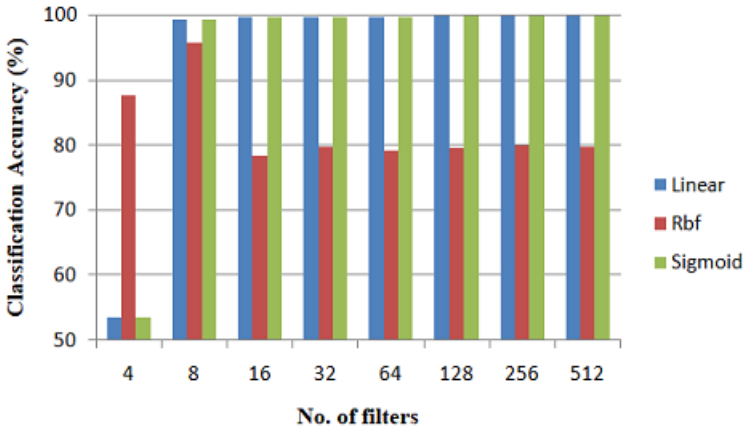

Figure 4.11: Analysis of kernel functions of SVM (Model 2 setup 1)

4. **Effect of the increase in Number of filters**

   The performance of this model increases till 8 filters. But then, as discussed in case of Model 1, the accuracy remains almost constant with the increase in number of filters after 8 filters. Figure 4.12 describes the effect of increase in number of filters on classification accuracy of model.

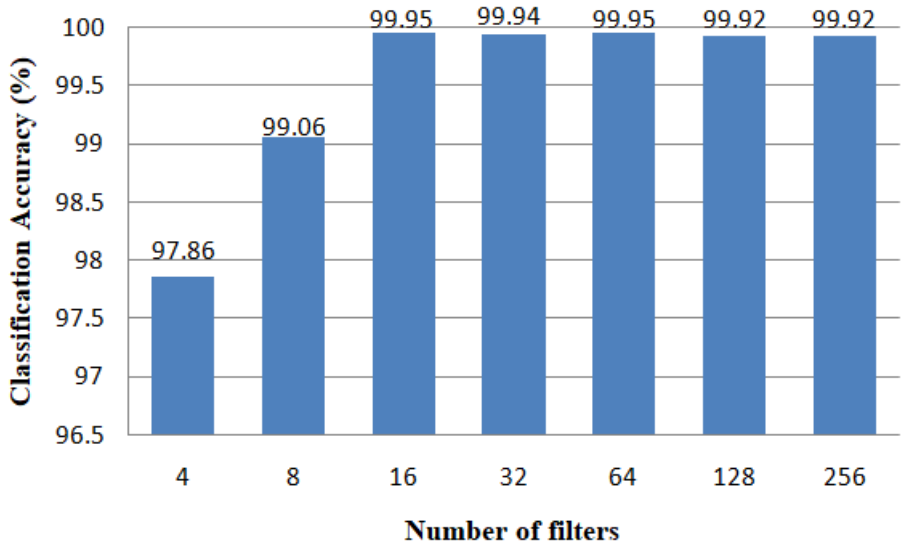

Figure 4.12: Effect of number of filters on performance (Model 2 setup 1)

Best Combination of parameters for Model 2: 4-layered CNN model and classification through SVM

Table 4.11 illustrates the best combination of parameters for proposed approach in which four-layered CNN is used and SVM is used for classification.

Table 4.11: Best combination of hyperparameters for Model 2

Hyperparameters	**Values**
Filter size(conv1)	9*9
No. of filters(conv1)	8
Pool size(maxpool1)	7*7
Filter size(conv2)	3*3
No. of filters(conv2)	8
Pool size(maxpool2)	3*3
Stride Length	1*1
Kernel function	Linear or Sigmoid

4.4.4 Comparative Analysis of Model 1 and Model 2 of the proposed approach

Model 1 of the proposed approach has two-layered CNN and Model 2 of the proposed approach has four-layered CNN. In both the models, classification is through SVM. The increase in number of layers has no considerable effect on classification accuracy of SVM. The classification accuracy in case of Model 1 of proposed aproach is 99.06% while

it is 99.29% in case of Model 2 of proposed approach. Infact, increasing the number of layers increase the runtime of model. So, two-layered CNN model has an advantage over four-layered CNN model in terms of runtime and accuracy.

4.5 Performance comparison of three approaches

Figure 4.13 illustrates the comparative analysis of three approaches i.e human-defined features based machine learning approach (MFCCs and SVM) (section 4.4.1), human-defined features based deep learning approach (MFCCs and CNN) (section 4.4.2) and the proposed off the shelf CNN features based approach (CNN and SVM) (section 3.3.1) for vehicle classification using acoustics.

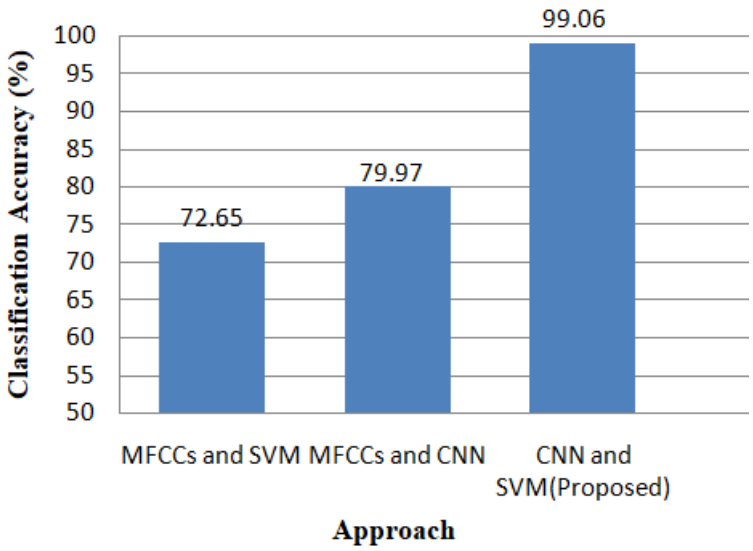

Figure 4.13: Comparative Analysis of three approaches

4.6 Feature Fusion

Feature Fusion is described in section 3.4. Number of features from intermediate layer of CNN vary with the change in number of filters and stride (Table 4.4). 13 Mel Frequency Cepstral Coefficients are extracted and combined with features from CNN. Following is the list of the results of experiments in case of feature fusion.

4.6.1 Effect of increase in stride length

As the stride increases, test accuracy of feature fusion increases. Figure 4.14 and Table 4.12 illustrate the effect of increase of stride length on classification accuracy of model.

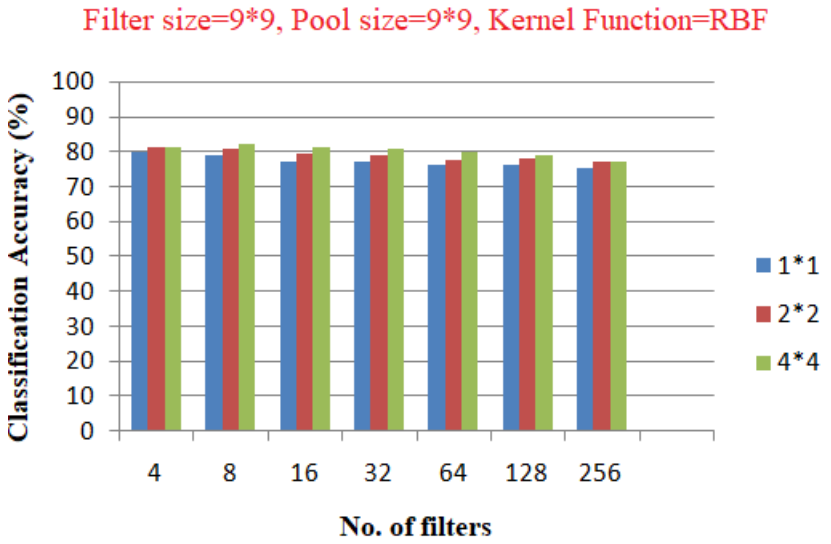

Figure 4.14: Effect of stride on classification accuracy (Feature Fusion)

Table 4.12: Effect of stride length on classification accuracy (Feature Fusion)

No. of filters	**Stride length**		
	1*1	2*2	4*4
4	79.80	81.16	81.52
8	79.12	80.69	82.15
16	77.24	79.70	81.26
32	77.24	79.12	80.79
64	76.12	77.82	79.78
128	76.10	78.12	78.90
256	75.78	77.11	78.90
512	75.23	77.01	77.59

4.6.2 Effect of Increase in number of filters

As the number of filters increases from 4 to 256, the accuracy of classification remains almost constant. The effect of increase in number of filters on classification accuracy of model is depicted in figure 4.15.

4.6.3 Effect of pool size in max-pooling layer

The pool size of 9*9 yields the best accuracy. Figure 4.16 describes the effect of pool size on classification accuracy of the model.

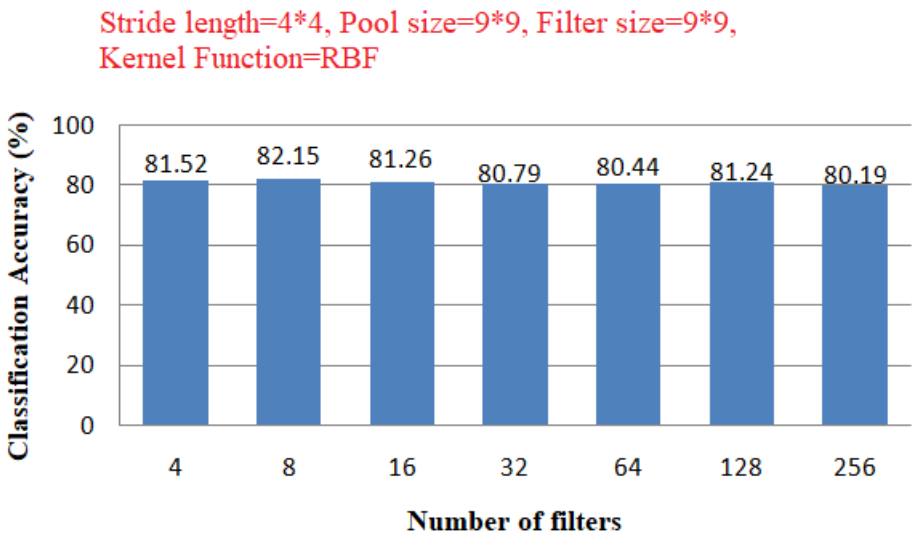

Figure 4.15: Effect of Number of filters on performance of model (Feature Fusion)

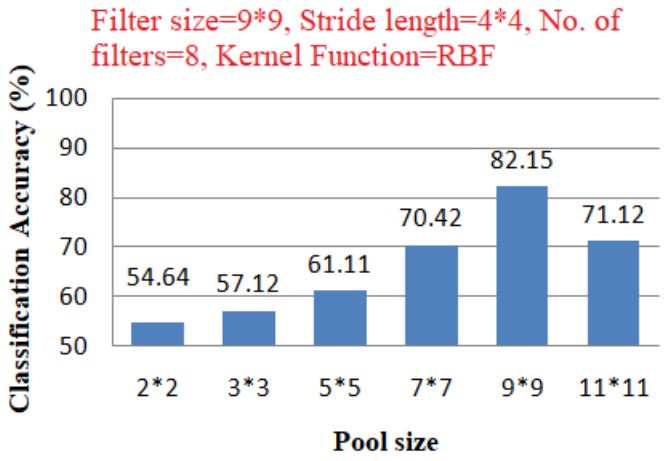

Figure 4.16: Effect of pool size on performance of model (Feature Fusion)

4.6.4 Effect of Size of filter in convolutional layer

The filter size of 9*9 gives the best results. Figure 4.17 shows the effect of the size of filter on the performance of model.

4.6.5 Analysis of Various kernel functions

RBF gives better results as compared to linear and sigmoid kernel functions. Table 4.13 and figure 4.18 illustrate the variation of performance of model with different kernels of SVM in case of feature fusion

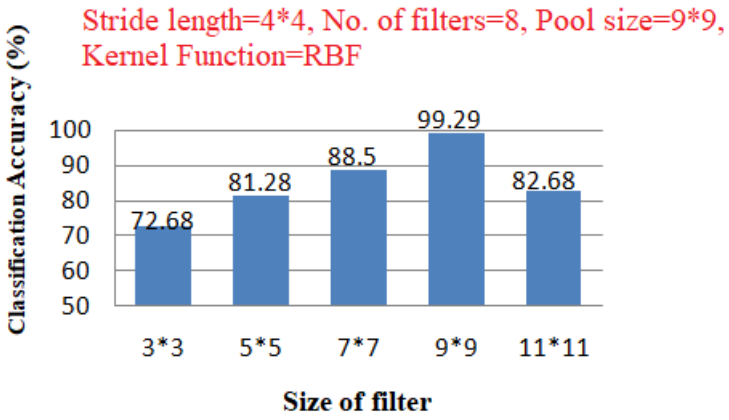

Figure 4.17: Effect of size of filter on classification accuracy (Feature Fusion)

Table 4.13: Analysis of kernel functions of SVM (Feature Fusion)

	Kernel function		
No. of filters	Linear	Rbf	Sigmoid
4	64.87	81.52	55.95
8	56.89	82.15	56.52
16	59.81	81.26	57.78
32	65.55	80.79	60.91
64	62.84	80.44	57.68
128	66.96	81.24	61.52
256	60.45	80.19	56.82

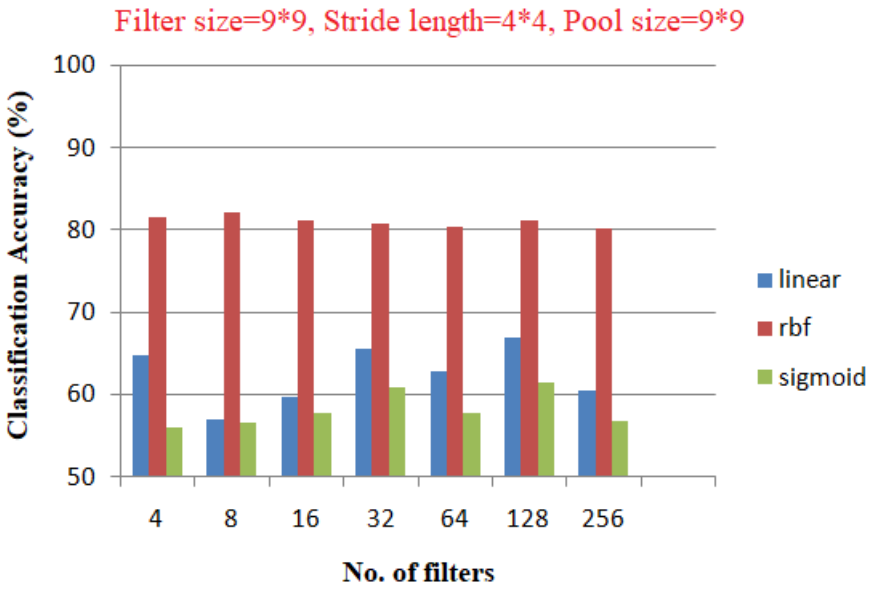

Figure 4.18: Analysis of kernel functions of SVM (Feature Fusion)

4.7 Analysis of Feature Fusion

We have fused the human-defined features i.e MFCCs with the features extracted using CNN. The fusion of features is carried out using the model 1 (section 3.3.1). The fused features are fed to SVM but this feature fusion does not give better results than the proposed approach (section 4.4.3) in which features extracted using CNN are fed to the SVM. The experiments are performed by varying various hyperparameters like filter size, number of filters, pool size, stride length so as to get the adequate combination of tuned hyperparameters which can give the best accuracy. The overall classification accuracy in case of feature fusion is 82.15% in case of 4 filters of size 9*9, pol size of 9*9 and stride length of 4*4 which is considerably less than the proposed approach (99.06%). So, the fusion of human-defined features and features extracted using CNN does not work better than the features extracted using CNN for vehicle classification using acoustics.

4.8 Summary

In this chapter, results of the proposed model for vehicle classification using acoustics and its comparison with two other approaches are discussed. Experiments are performed to determine adequate combination of various hyperparameters i.e. filter size, number of filters, pool size, stride length and kernel function. We have also described the results of feature fusion i.e. fusing off the shelf CNN features and human-defined features. Two models of the proposed approach are discussed- two-layered CNN model and four-layered CNN model. In both the models, SVM classifies the audio recordings into one of the vehicle categories. In human-defined features based machine learning approach (MFCCs and SVM), RBF gives the best accuracy of 72.65%. In human-defined features based deep learning approach (MFCCs and CNN), classification accuracy of 79.97% is obtained. The two-layered CNN model with SVM gives the best results (99.06%) for 8 filters each of dimension 9*9, stride length of 2*2, pool size of 9*9. The four-layered CNN model with SVM gives the best results (99.29%) for 8 filters each of dimension 9*9 in first convolutional layer, stride length of 1*1. For feature fusion, the best results (82.15%) are obtained with 4 filters each of dimension 9*9, stride length of 4*4, pool size of 9*9 in max pooling layer and Radial Basis kernel function.

So, the proposed off the shelf CNN features based approach for vehicle classification using acoustics outperforms the other approaches used for vehicle classification in this work. The combination of human-defined features and off the shelf CNN features fails to give better results than the proposed approach where the features extracted from CNN are employed for vehicle classification using SVM.

Chapter 5

Conclusion

Vehicle Classification is important for traffic modeling and in applications like automatic toll collection, parking, vehicle identification and surveillance etc. Various sensing techniques-both infrastructure based and infrastructureless techniques for recognition of vehicles' categories are costly, require high installation and maintenance costs and need high computational power. In this thesis, sounds of vehicles, collected through smartphones based acoustic sensors are employed for vehicle classification. A dataset consisting of 4789 recordings is used. We have proposed an off the shelf Convolutional Neural Network (CNN) based feature extraction approach. First the audio recordings of the vehicles undergo preprocessing. The audio signals are bifurcated into frames with frame length of 8192 and features are extracted using CNN. These features are given to Support Vector Machine (SVM) which classifies the audio signals into one of the public transport vehicles' categories- car, bus, aeroplane, train and three-wheeler. This approach is compared with the MFCCs based machine learning approach (SVM) and MFCCs based deep learning approach (CNN).

We have experimented with two models of the proposed approach- two-layered CNN and four-layered CNN model so as to determine the effect of the increase in number of layers on classification accuracy of the model. In both the models of proposed approach, CNN is used for feature extraction and SVM is used for classifying audio recordings into vehicle categories. Experiments are performed to determine the adequate combination of hyperparameters. So, the different number of filters, dimensions of filters, stride length, pooling size and different types of kernel functions are tested. The number of features extracted from CNN varies with the number of filters and stride length. Two-layered CNN model with SVM gives an accuracy of 99.06% in case of 8 filters each of size 9*9, stride length of 2*2, pool size of 9*9 and linear or sigmoid kernel function. The number of features extracted in this case was 32. Though the accuracy obtained in case

of four-layered CNN model with SVM is 99.29% which is not a considerable difference. Increasing the number of layers just increase the computational cost. So, two-layered CNN model is preferable. It outperforms the MFCCs based SVM approach (72.65%) and MFCCs based CNN approach (79.97%).

We have also explored if feature fusion can give good results than off the shelf features extracted from CNN. The features extracted using CNN are combined with the 13 MFCCs features from audio signals. These features are given to SVM which is trained to classify the vehicles. An accuracy of 82.15% is obtained with 8 filters each of dimension 9*9, pool size of 9*9, stride length of 4*4 and using RBF kernel function for SVM.

The results obtained have shown that the proposed approach is reliable in classifying vehicles using sound signals from vehicles. Using CNN for feature extraction overweighs the human-defined features for vehicle classification. Feature fusion does not help much in achieving higher accuracy as compared to employing off the shelf CNN features. We anticipate that the proposed approach using off the shelf CNN features for vehicle classification using acoustics can be employed in other acoustic domains by tuning various hyperaparameters.

5.1 Scope for future work

The results of the experiments in this thesis open up the ways to explore the network architecture further by testing more hyperparameters. The work in this thesis can be ameliorated and various other facets can be considered:

Cascading of deep learning architectures

Researchers have used recurrent neural networks [114] after number of convolutional layers and max-pooling layers [115] [116]. Similarly, other deep learning architectures such as Long Short Term Memory (LSTM) neural network [117] [118] can be cascaded after off the shelf CNN features [107].

Fusion of other types of features with off the shelf CNN features

We have used MFCCs for the vehicle classification in the other two approaches when compared with the proposed approaches. Other temporal and spectral features, wavelet features [119] or other features [120] can be used. These features can also be fused with off the shelf CNN features.

Transfer Learning

Various researchers have employed transfer learning for audio scene classification [121] [122] [123]. It is anticipated that the proposed approach can work well in other acoustic domains using transfer learning methodology and tuning various hyperparameters.

References

[1] B. Singh, N. Kapur, and P. Kaur, "Speech recognition with hidden markov model: A review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 3, 2012.

[2] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template-based continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1377–1390, 2007.

[3] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, IEEE, vol. 1, 1996, pp. 373–376.

[4] J. Liu and L. Xie, "Svm-based automatic classification of musical instruments," in *Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on*, IEEE, vol. 3, 2010, pp. 669–673.

[5] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[6] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on rms and zero-crossings," *IEEE Transactions on multimedia*, vol. 7, no. 1, pp. 155–166, 2005.

[7] Z. Ali, M. Alsulaiman, G. Muhammad, A. Al-nasheri, and A. Mahmood, "Clinical informatics: Mining of pathological data by acoustic analysis," in *Informatics, Health & Technology (ICIHT), International Conference on*, IEEE, 2017, pp. 1–8.

[8] L. Dickinson and N. H. Fletcher, "Acoustic detection of invisible damage in aircraft composite panels," *Applied Acoustics*, vol. 70, no. 1, pp. 110–119, 2009.

[9] R. O. Nielsen, "Acoustic detection of low flying aircraft," in *Technologies for Homeland Security, 2009. HST'09. IEEE Conference on*, IEEE, 2009, pp. 101–106.

[10] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, IEEE, 2005, 4–pp.

[11] J.-C. Wang, Y.-S. Lee, C.-H. Lin, E. Siahaan, and C.-H. 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.

[12] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "Dcase 2016 acoustic scene classification using convolutional neural networks," in *Proc. Workshop Detection Classif. Acoust. Scenes Events*, 2016, pp. 95–99.

[13] 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.

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

[15] J. Yu, H. Zhu, H. Han, Y. J. Chen, J. Yang, Y. Zhu, Z. Chen, G. Xue, and M. Li, "Senspeed: Sensing driving conditions to estimate vehicle speed in urban environments," *IEEE Transactions on Mobile Computing*, vol. 15, no. 1, pp. 202–216, 2016.

[16] R. Sen, P. Siriah, and B. Raman, "Roadsoundsense: Acoustic sensing based road congestion monitoring in developing regions," in *Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2011 8th Annual IEEE Communications Society Conference on*, IEEE, 2011, pp. 125–133.

[17] H. M. Ali and Z. S. Alwan, "Car accident detection and notification system using smartphone," *International Journal of Computer Science and Mobile Computing*, vol. 4, no. 4, pp. 620–35, 2015.

[18] F. Aloul, I. Zualkernan, R. Abu-Salma, H. Al-Ali, and M. Al-Merri, "Ibump: Smartphone application to detect car accidents," in *Industrial Automation, Information and Communications Technology (IAICT), 2014 International Conference on*, IEEE, 2014, pp. 52–56.

[19] M. P. Paulraj, A. H. Adom, S. Sundararaj, and N. B. A. Rahim, "Moving vehicle recognition and classification based on time domain approach," *Procedia Engineering*, vol. 53, pp. 405–410, 2013.

[20] V. Tyagi, S. Kalyanaraman, and R. Krishnapuram, "Vehicular traffic density state estimation based on cumulative road acoustics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1156–1166, 2012.

[21] G. De Angelis, A. De Angelis, V. Pasku, A. Moschitta, and P. Carbone, "A simple magnetic signature vehicles detection and classification system for smart cities," in *Systems Engineering (ISSE), 2016 IEEE International Symposium on*, IEEE, 2016, pp. 1–6.

[22] J. Gajda, R. Sroka, M. Stencel, A. Wajda, and T. Zeglen, "A vehicle classification based on inductive loop detectors," in *Instrumentation and Measurement Technology Conference, 2001. IMTC 2001. Proceedings of the 18th IEEE*, IEEE, vol. 1, 2001, pp. 460–464.

[23] O. Hasegawa and T. Kanade, "Type classification, color estimation, and specific target detection of moving targets on public streets," *Machine Vision and Applications*, vol. 16, no. 2, pp. 116–121, 2005.

[24] M. Simoncini, F. Sambo, L. Taccari, L. Bravi, S. Salti, and A. Lori, "Vehicle classification from low frequency gps data," in *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, IEEE, 2016, pp. 1159–1166.

[25] W. Ma, D. Xing, A. McKee, R. Bajwa, C. Flores, B. Fuller, and P. Varaiya, "A wireless accelerometer-based automatic vehicle classification prototype system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 1, pp. 104–111, 2014.

[26] B. Barbagli, G. Manes, R. Facchini, and A. Manes, "Acoustic sensor network for vehicle traffic monitoring," *VEHICULAR 2012*, p. 7, 2012.

[27] L. A. K. Luz Elena Y. Mimbela. (2000). Summary of vehicle detection and surveillance technologies used in intelligent transportation systems, [Online]. Available: `https://www.fhwa.dot.gov/ohim/tvtw/vdstits.pdf`.

[28] D. T. V. Mathew. (2014). Non-intrusive technologies, [Online]. Available: `http://nptel.ac.in/courses/105101008/downloads/cete_10.pdf`.

[29] ——, (2014). Intrusive technologies, [Online]. Available: `http://nptel.ac.in/courses/105101008/downloads/cete_09.pdf`.

[30] D. Vij and N. Aggarwal, "Smartphone based traffic state detection using acoustic analysis and crowdsourcing," *Applied Acoustics*, vol. 138, pp. 80–91, 2018.

[31] A. Kaur, N. Sood, N. Aggarwal, D. Vij, and B. Sachdeva, "Traffic state detection using smartphone based acoustic sensing," *Journal of Intelligent & Fuzzy Systems*, no. Preprint, pp. 1–8, 2017.

[32] R. Rekha and R. Karthika, "Fuzzy based traffic congestion detection & pattern analysis using inductive loop sensor," *International Journal of Scientific & Engineering Research*, vol. 4, no. 6, pp. 1149–1152, 2013.

[33] B. Krause, C. von Altrock, and M. Pozybill, "Intelligent highway by fuzzy logic: Congestion detection and traffic control on multi-lane roads with variable road signs," in *Fuzzy Systems, 1996., Proceedings of the Fifth IEEE International Conference on*, IEEE, vol. 3, 1996, pp. 1832–1837.

[34] S. Meta and M. G. Cinsdikici, "Vehicle-classification algorithm based on component analysis for single-loop inductive detector," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 6, pp. 2795–2805, 2010.

[35] Z.-X. Li, X.-M. Yang, and Z. Li, "Application of cement-based piezoelectric sensors for monitoring traffic flows," *Journal of transportation engineering*, vol. 132, no. 7, pp. 565–573, 2006.

[36] J. V. Chatigny, M. Thompson, P. F. Radice, and D. L. Halvorsen, *Traffic sensor having piezoelectric sensors which distinguish lanes*, US Patent 5,486,820, Jan. 1996.

[37] P. Burnos, J. Gajda, P. Piwowar, R. Sroka, M. Stencel, and T. Zeglen, "Measurements of road traffic parameters using inductive loops and piezoelectric sensors," 2007.

[38] R. Sroka, "Data fusion methods based on fuzzy measures in vehicle classification process," in *Instrumentation and Measurement Technology Conference, 2004. IMTC 04. Proceedings of the 21st IEEE*, IEEE, vol. 3, 2004, pp. 2234–2239.

[39] S. A. Rajab, A. S. Othman, and H. H. Refai, "Novel vehicle and motorcycle classification using single element piezoelectric sensor," in *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, IEEE, 2012, pp. 496–501.

[40] S. A. Rajab and H. H. Refai, "A single element piezoelectric sensor for vehicle classification using the ratio of track width to length," in *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, IEEE, 2014, pp. 1463–1468.

[41] S. A. Rajab, A. Mayeli, and H. H. Refai, "Vehicle classification and accurate speed calculation using multi-element piezoelectric sensor," in *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, IEEE, 2014, pp. 894–899.

[42] S. Kaewkamnerd, R. Pongthornseri, J. Chinrungrueng, and T. Silawan, "Automatic vehicle classification using wireless magnetic sensor," in *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2009. IDAACS 2009. IEEE International Workshop on*, IEEE, 2009, pp. 420–424.

[43] Y. Wang, "Joint random field model for all-weather moving vehicle detection," *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2491–2501, 2010.

[44] L.-W. Tsai, J.-W. Hsieh, and K.-C. Fan, "Vehicle detection using normalized color and edge map," *IEEE transactions on Image Processing*, vol. 16, no. 3, pp. 850–864, 2007.

[45] M. Vargas, J. M. Milla, S. L. Toral, and F. Barrero, "An enhanced background estimation algorithm for vehicle detection in urban traffic scenes," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 8, pp. 3694–3709, 2010.

[46] S. Paygude, Vyasvibha, C. Manisha, and I. Puneuniversity, "Vehicle detection and tracking using the optical flow and background subtraction," 2013.

[47] J.-C. Lai, S.-S. Huang, and C.-C. Tseng, "Image-based vehicle tracking and classification on the highway," in *Green Circuits and Systems (ICGCS), 2010 International Conference on*, IEEE, 2010, pp. 666–670.

[48] A. Tsuge, H. Takigawa, H. Osuga, H. Soma, and K. Morisaki, "Accident vehicle automatic detection system by image processing technology," in *Vehicle Navigation and Information Systems Conference, 1994. Proceedings., 1994*, IEEE, 1994, pp. 45–50.

[49] W. Zhang, Q. J. Wu, X. Yang, and X. Fang, "Multilevel framework to detect and handle vehicle occlusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 1, pp. 161–174, 2008.

[50] L. T. Ng, S. A. Suandi, and S. S. Teoh, "Vehicle classification using visual background extractor and multi-class support vector machines," in *The 8th International Conference on Robotic, Vision, Signal Processing & Power Applications*, Springer, 2014, pp. 221–227.

[51] L. Zhuo, L. Jiang, Z. Zhu, J. Li, J. Zhang, and H. Long, "Vehicle classification for large-scale traffic surveillance videos using convolutional neural networks," *Machine Vision and Applications*, vol. 28, no. 7, pp. 793–802, 2017.

[52] E. Walton, I. Theron, S. Gunawan, and L. Cai, "Moving vehicle range profiles measured using a noise radar," in *Antennas and Propagation Society International Symposium, 1997. IEEE., 1997 Digest*, IEEE, vol. 4, 1997, pp. 2597–2600.

[53] S. Park, J. P. Hwang, E. Kim, and H.-J. Kang, "Vehicle tracking using a microwave radar for situation awareness," *Control Engineering Practice*, vol. 18, no. 4, pp. 383–395, 2010, ISSN: 0967-0661. DOI: https://doi.org/10.1016/j.conengprac.2009.12.006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967066109002342.

[54] M. Cherniakov, R. R. Abdullah, P. Jancovic, and M. Salous, "Forward scattering micro sensor for vehicle classification," in *Radar Conference, 2005 IEEE International*, IEEE, 2005, pp. 184–189.

[55] I. Urazghildiiev, R. Ragnarsson, K. Wallin, A. Rydberg, P. Ridderstrom, and E. Ojefors, "A vehicle classification system based on microwave radar measurement of height profiles," 2002.

[56] E. Oudat, M. Mousa, and C. Claudel, "Vehicle detection and classification using passive infrared sensing," in *Mobile Ad Hoc and Sensor Systems (MASS), 2015 IEEE 12th International Conference on*, IEEE, 2015, pp. 443–444.

[57] S. Tropartz, E. Horber, and K. Gruner, "Experiences and results from vehicle classification using infrared overhead laser sensors at toll plazas in new york city," in *Intelligent Transportation Systems, 1999. Proceedings. 1999 IEEE/IEEJ/JSAI International Conference on*, IEEE, 1999, pp. 686–691.

[58] E. Odat, J. S. Shamma, and C. Claudel, "Vehicle classification and speed estimation using combined passive infrared/ultrasonic sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1593–1606, 2018.

[59] H. Maciejewski, J. Mazurkiewicz, K. Skowron, and T. Walkowiak, "Neural networks for vehicle recognition," in *Proceeding of the 6th International Conference on Microelectronics for Neural Networks, Evolutionary and Fuzzy Systems*, 1997, p. 5.

[60] H. Wu, M. Siegel, and P. Khosla, "Vehicle sound signature recognition by frequency vector principal component analysis," in *Instrumentation and Measurement Technology Conference, 1998. IMTC/98. Conference Proceedings. IEEE*, IEEE, vol. 1, 1998, pp. 429–434.

[61] M. E. Munich, "Bayesian subspace methods for acoustic signature recognition of vehicles," in *Signal Processing Conference, 2004 12th European*, IEEE, 2004, pp. 2107–2110.

[62] R. Sen, B. Raman, and P. Sharma, "Horn-ok-please," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*, ACM, 2010, pp. 137–150.

[63] J. George, A. Cyril, B. I. Koshy, and L. Mary, "Exploring sound signature for vehicle detection and classification using ann," *International Journal on Soft Computing*, vol. 4, no. 2, p. 29, 2013.

[64] J. George, L. Mary, and K. Riyas, "Vehicle detection and classification from acoustic signal using ann and knn," in *Control Communication and Computing (ICCC), 2013 International Conference on*, IEEE, 2013, pp. 436–439.

[65] M. Kandpal, V. K. Kakar, and G. Verma, "Classification of ground vehicles using acoustic signal processing and neural network classifier," in *Signal processing and communication (icsc), 2013 international conference on*, IEEE, 2013, pp. 512–518.

[66] M. R. Hamrick and R. M. Ingman, *Gps management system*, US Patent 9,734,698, Aug. 2017.

[67] A. S. Braunberger and B. M. Braunberger, *Absolute acceleration sensor for use within moving vehicles*, US Patent 7,239,953, Jul. 2007.

[68] E. D'Andrea and F. Marcelloni, "Detection of traffic congestion and incidents from gps trace analysis," *Expert Systems with Applications*, vol. 73, pp. 43–56, 2017.

[69] R. K. Balan, K. X. Nguyen, and L. Jiang, "Real-time trip information service for a large taxi fleet," in *Proceedings of the 9th international conference on Mobile systems, applications, and services*, ACM, 2011, pp. 99–112.

[70] J. Yoon, B. Noble, and M. Liu, "Surface street traffic estimation," in *Proceedings of the 5th international conference on Mobile systems, applications and services*, ACM, 2007, pp. 220–232.

[71] Z. Sun and X. J. Ban, "Vehicle classification using gps data," *Transportation Research Part C: Emerging Technologies*, vol. 37, pp. 102–117, 2013.

[72] S. Tao, V. Manolopoulos, S. Rodriguez Duenas, and A. Rusu, "Real-time urban traffic state estimation with a-gps mobile phones as probes," *Journal of Transportation Technologies*, vol. 2, no. 1, pp. 22–31, 2012.

[73] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich monitoring of road and traffic conditions using mobile smartphones," in *Proceedings of the 6th ACM conference on Embedded network sensor systems*, ACM, 2008, pp. 323–336.

[74] A. Mednis, G. Strazdins, R. Zviedris, G. Kanonirs, and L. Selavo, "Real time pothole detection using android smartphones with accelerometers," in *Distributed Computing in Sensor Systems and Workshops (DCOSS), 2011 International Conference on*, IEEE, 2011, pp. 1–6.

[75] R. Bhoraskar, N. Vankadhara, B. Raman, and P. Kulkarni, "Wolverine: Traffic and road condition estimation using smartphone sensors," in *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on*, IEEE, 2012, pp. 1–6.

[76] P. Singh, N. Juneja, and S. Kapoor, "Using mobile phone sensors to detect driving behavior," in *Proceedings of the 3rd ACM Symposium on Computing for Development*, ACM, 2013, p. 53.

[77] S. K. Banchhor and A. Khan, "Musical instrument recognition using zero crossing rate and short-time energy," *Musical Instrument*, vol. 1, no. 3, 2012.

[78] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern recognition letters*, vol. 22, no. 5, pp. 533–544, 2001.

[79] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on speech and audio processing*, vol. 10, no. 7, pp. 504–516, 2002.

[80] M. Jalil, F. A. Butt, and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in *Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2013 International Conference on*, IEEE, 2013, pp. 208–212.

[81] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence content classification using audio features," in *SETN*, Springer, 2006, pp. 502–507.

[82] T. Giannakopoulos, "Study and application of acoustic information for the detection of harmful content, and fusion with visual information," *Department of Informatics and Telecommunications, vol. PhD. University of Athens, Greece*, 2009.

[83] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A multi-class audio classification method with respect to violent content in movies using bayesian networks," in *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, IEEE, 2007, pp. 90–93.

[84] H. Misra, S. Ikbal, S. Sivadas, and H. Bourlard, "Multi-resolution spectral entropy feature for robust asr," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, IEEE, vol. 1, 2005, pp. I–253.

[85] B. Gajic and K. K. Paliwal, "Robust speech recognition in noisy environments based on subband spectral centroid histograms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 600–608, 2006.

[86] S. Lee, J. Kim, and I. Lee, "Speech/audio signal classification using spectral flux pattern recognition," in *Signal Processing Systems (SiPS), 2012 IEEE Workshop on*, IEEE, 2012, pp. 232–236.

[87] M. Kos, Z. KačIč, and D. Vlaj, "Acoustic classification and segmentation using modified spectral roll-off and variance-based features," *Digital Signal Processing*, vol. 23, no. 2, pp. 659–674, 2013.

[88] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, "Speech recognition using mfcc," in *International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July*, 2012, pp. 28–29.

[89] M. Bahoura and C. Pelletier, "New parameters for respiratory sound classification," in *Electrical and Computer Engineering, 2003. IEEE CCECE 2003. Canadian Conference on*, IEEE, vol. 3, 2003, pp. 1457–1460.

[90] 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.

[91] J. T. Foote, "Content-based retrieval of music and audio," in *Multimedia Storage and Archiving Systems II*, International Society for Optics and Photonics, vol. 3229, 1997, pp. 138–148.

[92] N. P. Desai, C. Lehman, B. Munson, and M. Wilson, "Supervised and unsupervised machine learning approaches to classifying chimpanzee vocalizations," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1786–1786, 2018.

[93] P. Lam, L. Wang, H. Y. Ngan, N. H. Yung, and A. G. Yeh, "Outlier detection in large-scale traffic data by naïve bayes method and gaussian mixture model method," *Electronic Imaging*, vol. 2017, no. 9, pp. 73–78, 2017.

[94] Y. Zou, G. Shi, H. Shi, and Y. Wang, "Image sequences based traffic incident detection for signaled intersections using hmm," in *Hybrid Intelligent Systems, 2009. HIS'09. Ninth International Conference on*, IEEE, vol. 1, 2009, pp. 257–261.

[95] A. Gregoriades and A. Chrystodoulides, "Extracting traffic safety knowledge from historical accident data," in *Adjunct Proceedings of the 14th International Conference on Location Based Services*, ETH Zurich, 2018, pp. 109–114.

[96] K. Song, F. Li, X. Hu, L. He, W. Niu, S. Lu, and T. Zhang, "Multi-mode energy management strategy for fuel cell electric vehicles based on driving pattern identification using learning vector quantization neural network algorithm," *Journal of Power Sources*, vol. 389, pp. 230–239, 2018.

[97] S. Nawrin, M. R. Rahman, and S. Akhter, "Exploreing k-means with internal validity indexes for data clustering in traffic management system," *International Journal of Advanced Computer Science and Applications*, 2017.

[98] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Multimedia and Expo, 2005. ICME 2005. IEEE International conference on*, IEEE, 2005, pp. 1306–1309.

[99] J.-H. Choi and J.-H. Chang, "On using acoustic environment classification for statistical model-based speech enhancement," *Speech Communication*, vol. 54, no. 3, pp. 477–490, 2012.

[100] D. S. Tayade and S. S. Gharde, "Audio-visual detection and classification of vehicle using multiclass svm: A review," *International Journal of Scientific Engineering and Technology Research*, vol. 3, no. 7, pp. 1106–1109, 2012.

[101] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, `http://www.deeplearningbook.org`.

[102] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," in *Neural Networks (IJCNN), 2017 International Joint Conference on*, IEEE, 2017, pp. 1547–1554.

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

[104] S. Kahl, T. Wilhelm-Stein, H. Hussein, H. Klinck, D. Kowerko, M. Ritter, and M. Eibl, "Large-scale bird sound classification using convolutional neural networks," *Working notes of CLEF*, 2017.

[105] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, IEEE, 2017, pp. 131–135.

[106] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," *arXiv preprint arXiv:1604.06338*, 2016.

[107] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of lstm and cnn," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 11–15.

[108] N. Aggarwal, D. Vij, and S. Soni, *Acoustic vehicular data.* Available at: `http://pudataset.puchd.ac.in:8080/jspui/handle/123456789/14` [accessed 10 April 2018], 2017.

[109] K. G. Pasi and S. R. Naik, "Effect of parameter variations on accuracy of convolutional neural network," in *Computing, Analytics and Security Trends (CAST), International Conference on*, IEEE, 2016, pp. 398–403.

[110] H. Phan, L. Hertel, M. Maass, P. Koch, and A. Mertins, "Label tree embeddings for acoustic scene classification," in *Proceedings of the 2016 ACM on Multimedia Conference*, ACM, 2016, pp. 486–490.

[111] J. Schroder, N. Moritz, J. Anemuller, S. Goetze, B. Kollmeier, J. Schrofffdfffdder, N. Moritz, J. Anemufffdfffdller, 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 (TASLP)*, vol. 25, no. 6, pp. 1304–1314, 2017.

[112] K. Yousaf, A. Iftikhar, and A. Javed, "Comparative analysis of automatic vehicle classification techniques: A survey," *International Journal of Image, Graphics and Signal Processing*, vol. 4, no. 9, p. 52, 2012.

[113] T. Sinha, B. Verma, and A. Haidar, "Optimization of convolutional neural network parameters for image classification," in *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*, IEEE, 2017, pp. 1–7.

[114] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[115] G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen, *et al.*, "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.

[116] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *Neural Networks (IJCNN), 2017 International Joint Conference on*, IEEE, 2017, pp. 3461–3466.

[117] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth annual conference of the international speech communication association*, 2014.

[118] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[119] C.-C. Lin, S.-H. Chen, T.-K. Truong, and Y. 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.

[120] V. Elaiyaraja and P. Sundaram, "Audio classification using support vector machines and independent component analysis," *Journal of Computer Applications (JCA)*, vol. 5, no. 1, 2012.

[121] A. Fritzler, S. Koitka, and C. M. Friedrich, "Recognizing bird species in audio files using transfer learning," *Working Notes of CLEF*, vol. 2017, 2017.

[122] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," *arXiv preprint arXiv:1703.09179*, 2017.

[123] A. Van Den Oord, S. Dieleman, and B. Schrauwen, "Transfer learning by supervised pre-training for audio-based music classification," in *Conference of the International Society for Music Information Retrieval (ISMIR 2014)*, 2014.