KALID ZAMAN CLASSIFICATION OF SPEECH SIGNALS FOR HEARIN AID DEVICES USING CONVOLUTIOANL NEURAL NETWORKS NEU 2018

CLASSIFICATION OF SPEECH SIGNALS FOR HEARING AID DEVICES USING CONVOLUTIONAL NEURAL NETWORKS

A THESIS SUBMITTED TO THE GRADUATE SCHOOL OF APPLIED SCIENCES OF NEAR EAST UNIVERSITY

By
KHALID ZAMAN

In Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering

NICOSIA, 2019

CLASSIFICATION OF SPEECH SIGNALS FOR HEARING AID DEVICES USING CONVOLUTIONAL NEURAL NETWORKS

THESIS SUBMITTED TO THE
GRADUATE SCHOOL OF APPLIED SCIENCES
OF
NEAR EAST UNIVERSITY

By
Khalid Zaman

In Partial Fulfillment of the Requirements for the
Degree of Master of Science
in
Computer Engineering

NICOSIA, 2019

Khalid Zaman: CLASSIFICTION OF SPEECH SIGNALS FOR HEARING AID DEVICES USING CONVOLUTIONAL NEURAL NETWORKS

Approval of Director of Graduate School of Applied Sciences

Prof. Dr. Nadire ÇAVUŞ

We certify this thesis is satisfactory for the award of the degree of Master of Science in Computer Engineering

Examining Committee in Charge:

Prof. Dr. Rahib Abiyev	Committee Chairman, Department of Computer Engineering, NEU
Assoc. Prof. Dr. Kamil Dimililer	Department of Electrical and Electronics Engineering, NEU
Assoc. Prof. Dr. Melike Şah Direkoğlu	Supervisor, Department of Computer Engineering, NEU
Assist. Prof. Dr. Cem Direkoğlu	Co-supervisor, Department of Electrical and Electronics Engineering, METU NCC

I certify that this research work entitled “Classification of Harmful Noise for Hearing Aid Devices Using Convolutional Neural Networks” is my own work. No portion of the work presented in this research report has been submitted in support of another award or qualification either at this institution or elsewhere. Where material has been used from other sources it has been properly acknowledged / referred. If any part of this project is proved to be copied or found to be a report of some other, I will stand by the consequences.

Name, Last Name:

Signature:

Date

ACKNOWLEDGMENT

All praise and thanks is due to Almighty Allah, the Lord of mankind and that exists, for His blessings, benevolence and guidance at every stage of our life.

I am deeply grateful to my supervisor Assoc.Prof Dr Melike Şah Direkoğlu, for her guidance, support and patience. She has been invaluable source of knowledge and has certainly helped to inspire many of the ideas expressed in this thesis.

I would like to show my gratitude to my co-supervisor Assist Prof Dr Cem Direkoğlu for spending time and reading my thesis report and providing useful suggestions about this thesis.

I would wish to express my sincere regards to the Chairman of the Department of Computer Engineering Prof Dr Rahib Abiyev for his open hearted encouragement and good wishes.

Our words will fail to express our deepest heartfelt thanks to our families, especially my parents, and my dear uncle Mir Jehan for all what they did, and still doing to help us be at this position and for their Continuous support and encouragement.

ABSTRACT

Hearing loss is a major problem where many people in the world suffer from this disease. Normally hearing loss is occurred when someone loss the ability of hearing,which may occur in any part of the ear. Therefore, hearing aid devices are very important for impaired people suffering from hearing loss. There are different types of hearing aid devices to enhance the speech signals for hearing impaired people. In early years, people use analog hearing aid devices which work on the principle of amplifier and that devices cannot be adapoted different people. As a result, special solutions are required to be adapted by different individuals. To overcome this problem, digital hearing aid devices are adopted which are more flexible,can be programmedaccording to a specific frequency and able to adapt different hearing loss conditions compared to analog hearing aid devices. Generally these device use low-pass filter, adaptive filters and spectral analysis to remove harmful noise signal and amplify the incoming speech signal. However, in some cases, digital hearing aid devices cannot remove harmful noises, which may damage the ear. Therefore, this research focuses on classification of speech signals as harmful and unharmful using Convolution Neural Networks. In particular, first, we add different types of noises to the speech signal such as white noise, jet aircrafts noise, storm noise, fixes frequency noise. Then, we try to remove these noises using different speech filters and analyze the speech. We observe that for some noise types, speech cannot be cleaned properly. To overcome this, we apply Convolution Neural Networks (CNN) to classify

speech signals as harmful and unharmful in order to detect harmful speeches in digital hearing aid devices.

Keywords: **Hearing aid devices; Convolutional Neural Networks; fixed filters; adaptive filters**

ÖZET

İşitme kaybı, dünyadaki birçok insanın bu hastalıktan muzdarip olduğu önemli bir sorundur. Normal olarak, bir kişi kulağın herhangi bir yerinde oluşabilecek duyma yeteneğini kaybettiğinde işitme kaybı meydana gelir. Bu nedenle, işitme cihazı cihazları işitme kaybından muzdarip insanlar için çok önemlidir. İşitme engelli kişilerin konuşma sinyallerini geliştirmek için farklı işitme cihazı tipleri vardır. İlk yıllarda insanlar, amplifikatör prensibi üzerinde çalışan ve cihazların farklı kişilere adapte edilemeyecekleri analog işitme cihazı kullanıyorlar. Sonuç olarak, farklı kişiler tarafından uyarlanması için özel çözümler gerekir. Bu sorunun üstesinden gelmek için, daha esnek olan, belirli bir frekansa göre programlanabilen ve farklı işitme kaybı koşullarını analog işitme cihazlarına göre uyarlayabilen dijital işitme cihazı cihazları benimsenmiştir. Genellikle bu cihaz, zararlı parazit sinyalini ortadan kaldırmak ve gelen konuşma sinyalini güçlendirmek için alçak geçirgen filtre, uyarlamalı filtreler ve spektral analiz kullanır. Bununla birlikte, bazı durumlarda, dijital işitme cihazı, kulağa zarar verebilecek zararlı sesleri çıkaramaz. Bu nedenle, bu araştırma, konuşma sinyallerinin Convolution Sinir Ağları kullanılarak zararlı ve zararsız olarak sınıflandırılmasına odaklanmaktadır. Özellikle, önce beyaz gürültü, jet uçakları gürültüsü, fırtına gürültüsü,

frekans gürültüsünü gideren konuşma sinyaline farklı ses türleri ekleriz. Ardından, farklı konuşma filtreleri kullanarak bu sesleri gidermeye ve konuşmayı analiz etmeye çalışıyoruz. Bazı gürültü türleri için konuşmanın düzgün bir şekilde temizlenemediğini gözlemledik. Bunun üstesinden gelmek için, dijital işitme cihazlarındaki zararlı konuşmaları tespit etmek için konuşma sinyallerini zararlı ve zararlı olarak sınıflandırmak için Convolution Neural Networks'ü (CNN) kullanıyoruz.

Anahtar Kelimeler: İşitme cihazı; Dönüşümlü Yapay Sinir Ağları; sabit filtreler; uyarlanabilir filtreler

TABLE OF CONTENTS

ACKNOWLEDGEMENT ii

ABSTRACT iii

ÖZET iv

TABLE OF CONTENTS v

LIST OF FIGURES viii

LIST OF TABLES xi

ABBREVIATION xii

CHAPTER 1: INTRODUCTION

1.1.Intoduction 1

1.2. Objective 4

CHAPTER 2: LITERATURE REVIEW

2.1 How Human Ear Work 5

2.1.1 Outer ear 6

2.1.2 Middle ear 6

2.1.3 Inner ear 7

2.2 Normal Characteristics of Human Ear 8

2.3 Hearing Loss 9

2.3.1. Sensori-Neural hearing loss 9

2.3.2. Conductive hearing loss. 9

2.3.3. Conductive hearing loss. 9

2.4 Commercail Hearing Aids. 10

2.5 Factor of Hearing Loss. 10

2.6 Classification of Hearing Loss 10

2.7 Some Statistic About the Impaird People 11

2.8 Hearing Aid 12
2.8.1 Analog hearing aid device 12
2.8.2 Digital hearing aid device 13
2.9 Noise 14
2.9.1 Road traffic noises 14
2.9.2 Babble noise 14
2.9.3 External noise 14
2.9.4 Internal noise 15
2.10 Noise Cancelation Techniques 15
2.10.1 Noise cancelation using fixed filter 15
2.10.2 Noise cancelation usingadaptive filter 16
2.13 Spectral Analysis Tecnique 17

CHAPTER 3: **APPLYING DIGITAL FILTERS FOR NOISE REMOVAL**
3.1 Noise 19
3.1 Lowpass Filter 20
3.2 Aaptive Filter 22
3.3 Sepctral Analysis 25
3.4 Comments 26

CHAPTER 4: SPEECH SIGNALS AND CONVOLUTIONAL NEURAL NETWORKS
4.1 Speech Signals 27
4.1.1 Conversion of Speech Signals to Spectrogram Images 28
3.2 Convolutional Neural Networks 29
4.2.1 Convolutional layer 30
4.2.2 Max pooling layer: 31
4.2.3 Fully connected layer: 32

4.2.4 The Rectified linear unit 33
4.2.5 Soft max activation 34
4.2.6 Parameters intailization 34
4.2.7 Optimization 35
4.2.8 Regularization: 35
4.3 Speech Processing and Related Work 35

CHAPTER 5: CLASSIFICATION OF HARMFUL SPEECH SIGNALS USING CONVOLUTIONAL NEURAL NETWOKS

5.1 System Architecture 37
5.2 Speech and Noise Signal Spectrograms 38
5.3 Input spectrogram images to Convolutional Neural Network 40
5.4 Define Convolutional Neural Network Architecture 42
5.5 Train Network Using Training Data to Classify Validation Images and Compute Accuracy of Speech and High Frequency Fixed Noise 43
5.5.1CNN using one layer 43
5.5.2 CNN using two layer 45
5.5.3 CNN using three layer network 47

CHAPTER 6: EVOLUTIONS

6.1 Results - Data from the Same Dataset 50
6.2 Results - Data from Unseen Dataset 56
6.3 Results – Testing CNN network with different noise types 60
6.4 Confusion Matrix 65
6.5 Signals to Noise Ratio(SNR) 67
6.6Epoch Table 68

CHAPTER 6: CONCLUSION 69
REFERENCES 71
APENDIX 74

LIST OF FIGURES

Figure 1.1: Block diagram of noise reduction system 3

Figure 1.2: Block diagram of classification of noise by 3

Figure 2.1: Human ear structure 5

Figure 2.2: Outer ear structure 6

Figure 2.3: Middle ear structure. 7

Figure 2.4: The iner ear structure 8

Figure 2.5: Audible range of normal human ear. 8

Figure 2.6: Graph of impaired people in world. 11

Figure 2.7: Block diagram of simple analog hearing aid device 12

Figure 2.8: Block diagram of digital hearing aid device 13

Figure 2.9: Adaptive filter 17

Figure 3.1: Noise types 19

Figure 3.2: Flowchart of lowpass butter filter 20

Figure 3.3: Noise cancellation using 6^{th} order lowpass filter with 0.6 cutoff frequency 21

Figure 3.4: Noise cancellation using 6^{th} order lowpass filter with 0.6 cut off frequency 22

Figure 3.5: Flowchart of adaptive filter in noise cancellation 23

Figure 3.6: Noise cancellation using adaptive filter with step size 0.001 23

Figure 3.7: Noise Cancellation using adaptive filter with step size 0.11 24

Figure 3.8: Noise cancellation using spectral analysis 25

Figure 3.9: Gaussian noise cancellation using lowpass filter 26
Figure 3.10: High frequency noise cancellation using adaptive filter 26
Figure 4.1: Speech signal 27
Figure 4.2: Spectrogram of a car horn 29
Figure 4.3: General framework of CCN 30
Figure 4.4: 2-Dcross-correlation of CNN 31
Figure 4.5: Max pooling 32
Figure 4.6: Illustration of the fully connected structure 33
Figure 5.1: Block diagram of the CNN classification of speech and noise 37
Figure 5.2: Spectrogram of the speech signal 38
Figure 5.3: Spectrogram of the jet aircrafts noise added to speech signal 39
Figure 5.4: Spectrogram of the high frequency fixed noise added to speech signal 39
Figure 5.5: Spectrogram gray scale image of the speech and white noise 40
Figure 5.6: Spectrogram gray scale images of the speech and High frequency noise 41
Figure 5.7: Spectrogram gray scale images of the speech and storm noise 41
Figure 5.8: 28 by 28 pixel input image to CNNs network 42
Figure 5.9: Block diagram of CNN classification of speech and noise using one layers.....44
Figure 5.10: Validation accuracy of the speech and high frequency noise using one layer.....45
Figure 5.11: Block diagram of CNN classification of speech and noise using two layers 46
Figure 5.12:Validation accuracy of speech and fixed high frequency noise with two layer.....47
Figure 5.13: Block diagram of CNN classification of speech and noise using three layers 48
Figure 5.14: Validation accuracy of speech and high frequency noise with three layers 49
Figure 6.1: Validation and test values of speech and noise signal (one convolutional layer).....51
Figure 6.2: Validation and test values of speech and noise signals (two convolution layers) 52
Figure 6.3: Validation and test values of speech and noise signal (three convolutional layers)..53
Figure 6.4: Validation and test values of speech and noise signal (one convolutional layer).....54
Figure 6.5: Validation and test values of speech and noise signals (two convolution layers) 55
Figure 6.6: Validation and test values of speech and noise signal (three convolutional layers)..56
Figure 6.7: Validation and test values of speech and noise signal(one convolutional layer).....58
Figure 6.8: Validation and test values of speech and noise signal(two convolutional layer) 59
Figure 6.9: Validation and test values of speech and noise signa(three convolutional layer).....60

Figure 6.10: Validation and test values of speech and noise signal 61
Figure 6.11: Validation and test values of speech and noise signal 62
Figure 6.12: Validation and test values of speech and noise signal 63
Figure 6.13: Validation and test values of speech and noise signal 64
Figure 6.14: Validation and test values of speech and noise signal 65
Figure 6.15: Confusion matrix 66
Figure 6.16: Confusion matrix 67

LIST OF TABLES

Table 1.1: Classification of Hearing Loss 11

Table 6.1: Validation and test values of speech and noise signal (one convolutional layer) 51

Table 6.2: Validation and test values of speech and noise signal (two convolutional layer) 52

Table 6.3: Validation and test values of speech and noise signal (three convolutional layers) .. 53

Table 6.4: Validation and test values of speech and noise signal (one convolutional layers)..... 54

Table 6.5: Validation and test values of speech and noise signal (two convolutional layers) 55

Table 6.6: Validation and test values of speech and noise signal (three convolutional layers) .. 56

Table 6.7: Validation and test values of speech and noise signal (one convolutional layers)..... 57

Table 6.8: Validation and test values of speech and noise signal (two convolutional layers) 58

Table 6.9: Validation and test values of speech and noise signal (three convolutional layers) .. 59

Table 6.10: Validation and test values of speech and noise signal..... 61

Table 6.11: Validation and test values of speech and noise signal..... 62

Table 6.12: Validation and test values of speech and noise signal..... 63

Table 6.13: Validation and test values of speech and noise signal..... 64

Table 6.14: Validation and test values of speech and noise signal..... 65

Table 6.15: Different epoch for fixed high frequency noise 68

LIST OF ABBREVIATIONS

CNN: Convolutional Nueral Network

LMS: Least Mean Squre

CHAPTER 1
INTRODUCTION

1.1Introdution

Hearing loss is a major problem, where many people suffer worldwide. Normally hearing loss is occurred when someone loss the ability of hearing which may occur in any part of the ear. There are three parts of the human ear: Outer ear, middle ear and inner ear. The hearing loss problem occurs in any parts of the human ear. The factor that can cause the hearing loss problem are: age, usage of drugs, birth before the natural time diseases during pregnancy, the level of sound (intense sound), inherence, injuries of head and ear, diseases or illness etc.

Hearing loss can be classified as: Mild, moderate, severe and profound. According to (Halawani et al., 2013) mild people are not able to understand normal speech while people suffering from moderate hearing loss not well enough understand the loud speech. People with severe hearing loss can normally understand amplified speech and the people suffering are not profound able to well understand the amplified speech.

There are difference types of hearing aid device available to enhance the speech signals for the impaired peoples. First time people use the analog hearing aid device which just amplifies the speech signals. Before 1996 the majority of hearing aids were analog which work is an amplifier, not provided any kind of noise cancellation techniques and mechanisms. For the hearing impaired people analog hearing aid provided a generalized solution (Ngo, 2011). analog aid used small battery, microphone, speaker and simple electronic circuit which contain transistor to amplify and modify the sound which coming from outside and forward to inner ear. Analog hearing aid cannot difference between the speech and the noise signal, amplifying and modify both speech and noise. First digital hearing aid was introduced in 1996 which used digital signal processing to implement advance signal processing algorithm techniques. In United States in 2005 93% hearing aids sold that worked on digital signal processing technology (Ngo, 2011).The digital hearing aid design to know about the speech signal and noise. The digital hearing aids provided noise cancellation techniques in noisy environment

using advance degree of digital signal processing approach. For the hearing impaired people, analog hearing aid devices provide a generalized solution. It is stated that every individuals hearing characteristic is different and therefore for every individual there is a need to develop or design a special solution device according to the hearing impairment. For special solution of individual patients, they use a digital hearing aid which is more flexible and have ability to adjust the program of the hearing aid device according to different conditions of the individual compared to unchanging program of analog hearing kids. The digital hearing aids can be programmed to match the patients hearing loss individually according to a specific frequency. The aids are programmed using the human audiogram. Digital hearing aid can be work with a very low power battery, approximately in mW (Ngo, 2011). In a concise manner digital hearing aid enhance the clarity of sound with less circuit noise, faster processing of sound using digital signal processing, reduces incoming noise signal better than an analog hearing aid.

In this work use noise reduction system using digital filters and apply convolutional neural networks (CNN) for classification of harmful and safe speech signals for the human ear. The block diagram of noise reduction system that is implemented in Matlab is shown in Figure 1.1. First the incoming speech signal is passed through the noise reduction system to filter the speech signal from the noise and then output the filtered speech signal. In noise reduction system there are different types of noise cancellation techniques to remove noise from the speech signals. In noise cancellation techniques, fixed filters, adaptive filters and spectral subtraction techniques are used in this thesis. In fixed filters, we apply bandstop, bandpass, highpass and lowpass filters. Bandstop passes low frequencies and stops high frequencies. Highpass filters are used to remove low frequencies. Bandstop filters are used to remove a certain range of frequency. Finally, bandpass filters are used to pass a certain range of frequencies. Alternatively, adaptive filters can be utilized to remove background noise from the speech signals. Background noise changes by time to time (Halawani et al., 2013). Adaptive filter is able to adjust itself according to the background noise. Spectral subtraction technique is the earliest approach for the noise cancellation. In this technique, speech signals are recovered from the subtraction of estimated noisy signal spectrum which is received from noisy signal spectrum.

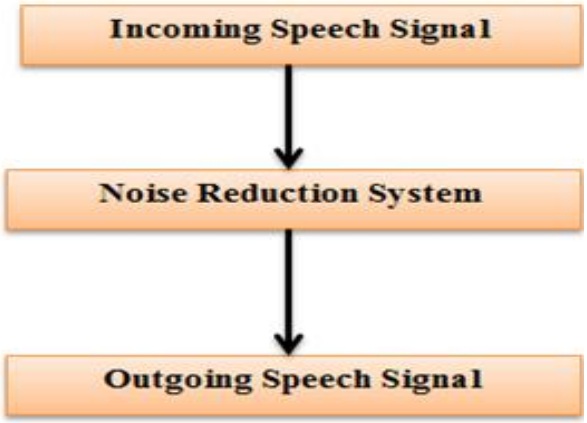


Figure 1.1: Block diagram of noise reduction system

Here we apply CNN algorithms to classify the speech and noise signals. In this technique, first speech and the added noise are converted to spectrogram images. Then CNN is applied to classify clean speech and noisy speech which is shown in Figure 1.2.

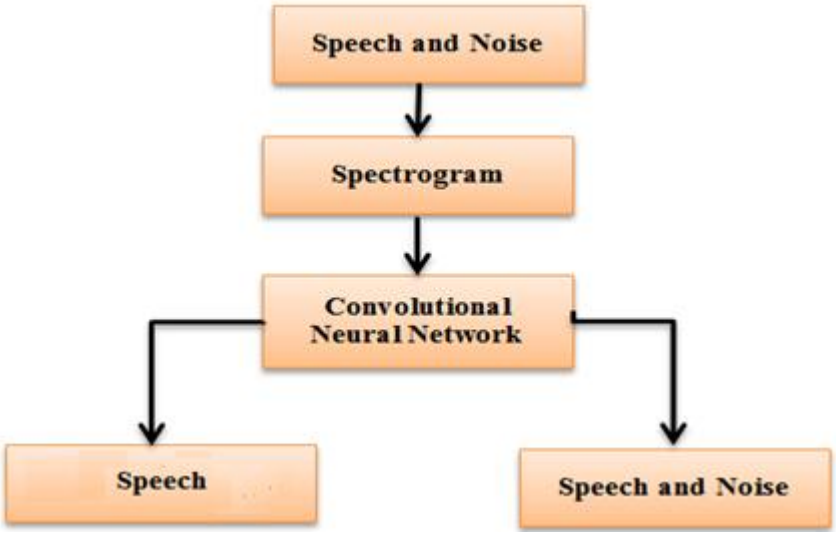


Figure 1.2: Block diagram of classification of noise by CNN

1.2 Objectives

The aim and objectives of the thesis are deeply study the speech signal and different kinds of noises and their removal techniques as well as classification of harmful and clean speech signals using CNN. The following are the different objectives of the thesis:

i. To study and analyze speech signals.
ii. To study and analyze different kinds of noises.
iii. To study and analyze convolution neural networks
iv. To study and understand the different techniques used to remove the noise.
v. To study and understand the how CNN can be used to classify speech and noise signals.
vi. To implement these techniques in Matlab in order to see the behaviour of signal and noise.

CHAPTER 2
LITERATURE REVIEW

2.1 How Human Ear Work

Human ear works to hear the sound which come from outside to the ear. The sound wave from outside to the inner ear is shown in the Figure 2.1. Sound waves come from outside enter to the ear canal cause to vibrate the eardrum (tympanic membrane). The eardrum passes the sound wave to vibrate the small bones of the middle ear (Wikipedia 2015). The middle ear leads the sound wave into inner ear. The cochlea inside the inner ear converts the mechanical sound vibration into electrical signal (pulse) or nerve impulses, and then sends electrical signals into brain (temporal lobe) through auditory nerve. The ear also concerned with sense of balancing.

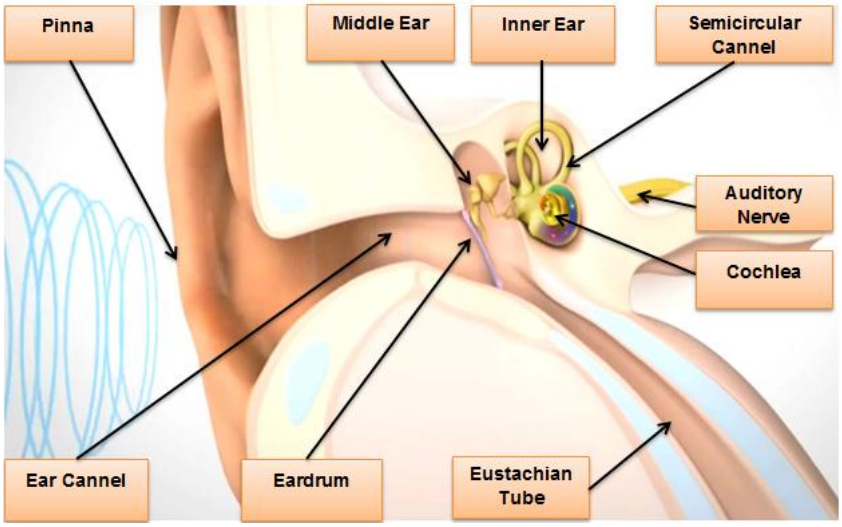


Figure 2.1: Human ear structure

2.1.1 Outer ear

Outer ear's task is to collect the sound wave and transmit to the middle ear which is shown in Figure 2.2. First, outer ear collects the sound wave through the ear pinna and then sound wave is passed through external auditory cannel to vibrate the eardrum (tympanic membrane) (nidcd, 2018). In the outer ear the sound wave use the free space medium for the transmission.

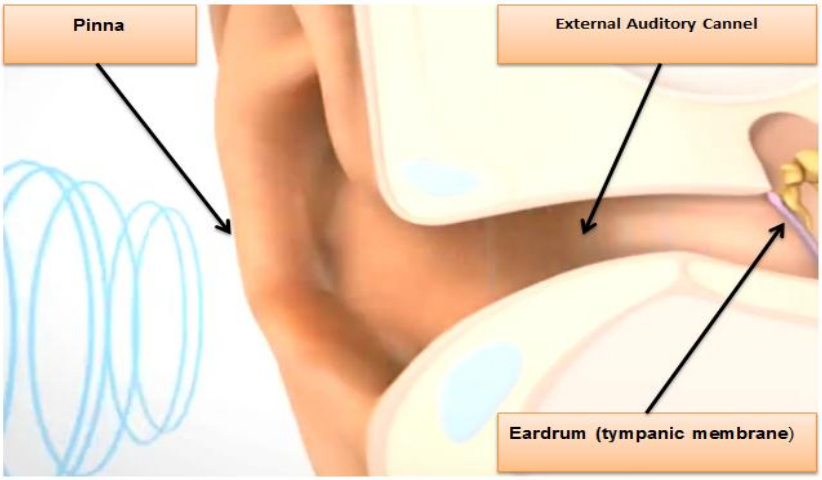


Figure 2.2: Outer ear structure

2.1.2 Middle ear

Middle ear sends the sound wave from the outer part of the ear to the inner part of the ear which is shown in Figure 2.3. Middle ear has three small and tiny bones (malleus, incus, and stapes). These three bones are responsible to lead or transmit the sound waves from eardrum to the inner part of the ear. These three bones also perform the work of amplification of the sound wave (nidcd, 2018).Eustachian tube which connects the middle part of the ear to the throat (pharynx). Its work is to equalize the pressure. It equalizes the pressure on the inner side

of the eardrum to the external pressure. Inside the middle ear sound waves use the solid medium for the transmission.

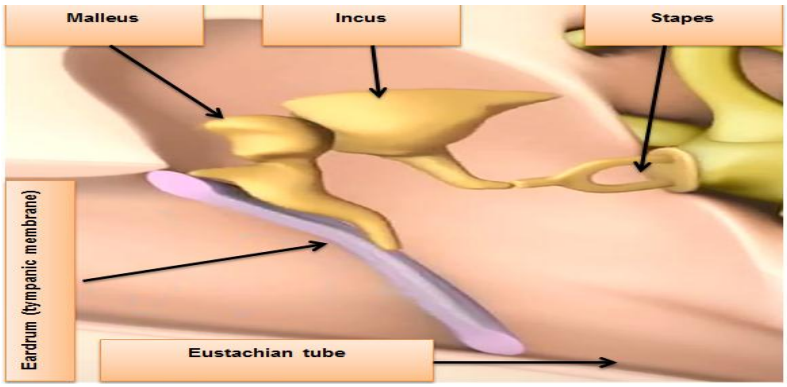
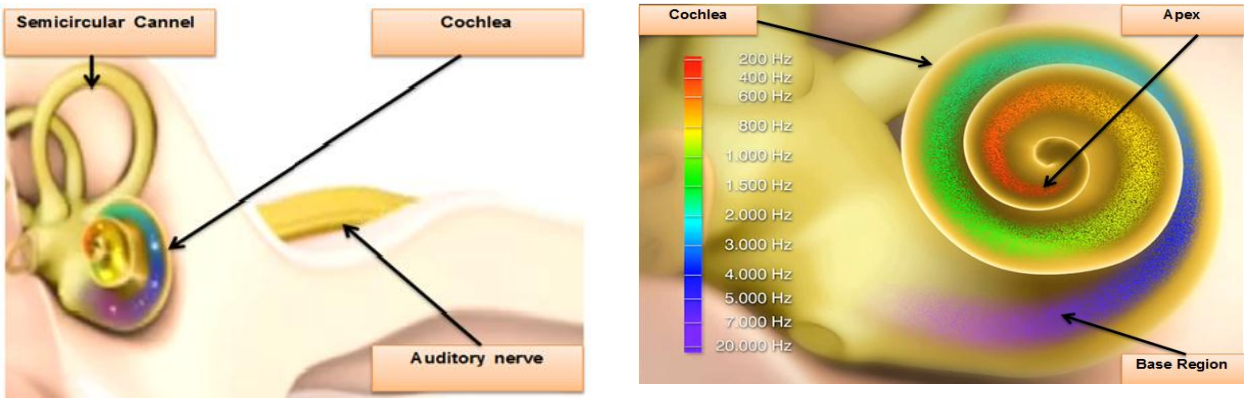


Figure 2.3: Middle ear structure

2.1.3 Inner ear

The inner ear transmits the sound wave from the middle part of the ear to the brain in shown in Figure 2.4. Inner ear translates the sound wave to electrical signals and then sends to the human brain. In the inner ear sound waves use the liquid medium for the transmission. Human brain only works on electrical pluses. Cochlea plays a key role to convert sound vibration into complex electrical signals. Then it is passed to brain by auditory nerve. Cochlea round like a coil or a snail shape fills with fluid. Inside the cochlea there are hair cells which have responsibility for detection of different frequencies. Hair cells allow perceiving the entire spectrum of sound. Hair cells inside cochlea arrange in order. Apex is responsible to manage low frequencies and base region is responsible for high frequencies (nidcd, 2018). Semicircular cannel is the part of inner ear responsible for balancing. Inside the inner ear sound waves use the liquid medium for the transmission.



(a) inner ear (b) The cocheala structure

Figure 2.4 : The Inner ear structure

2.2 Normal Characteristics of Human Ear

The normal characteristics of human ear are to hear the sound in range of frequency from 20 Hz to 20000Hz. The human ear is so sensible in frequency range of 1000 Hz to 5000 Hz (Halawani et al., 2013). The loudness of the sound is called intensity of the sound wave is the importance property of sound wave. Conventionally the intensity of sound wave measures with scale of decibel. The audible sound pressures which can human hear is range from 0db-120db which shows in the Figure 2.5.

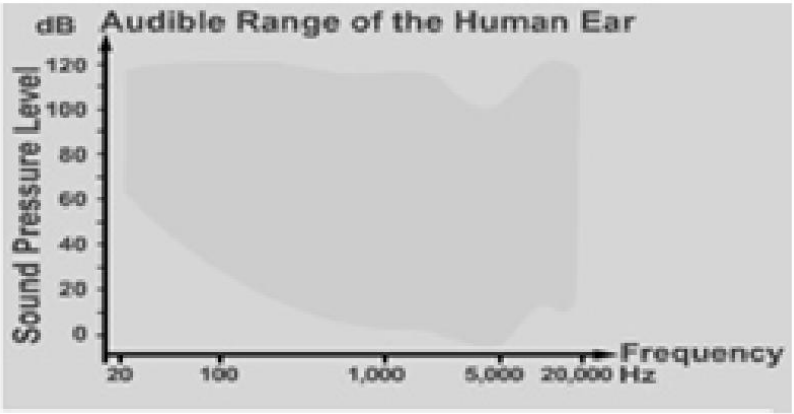


Figure 2.5: Normal human ear audible range

2.3 Hearing Loss

When someone loss the ability of hearingwhich may occurs in different parts of the human ear. The hearing loss occurs when someone have a problem in the inner ear, middle ear and outer ear. The types of hearing loss of a human are conductive, sensori-neural, and mixed hearing loss.

2.3.1 Sensori-Neural hearing loss

When problem occurred in cochlea of the human inner part of the ear is called sensori-neural hearing loss (Ngo, 2011).These types of loss are more severe than other ones. It ranges from mild to profound. This type person hears only a few frequencies than the normal person.

2.3.2 Conductive hearing loss

It is occurred when there is a problem in a conduction pathway to transfer sound to the inner ear (Ngo, 2011).Usually problem is in both inner part and outer part of the ears. This type of loss is inefficient transfer of sound occurred from the outside to the inner ear.

2.3.3 Mixed hearing loss

When both conductive and Sensori-Neural hearing loss is occurred in an ear is called mixed hearing loss (Ngo, 2011). In mixed hearing loss, the problem is all parts of the ear (outer, middle and inner ear).

2.4 Commercial Hearing Aids

Different types of commercial hearing aids devices are available in different shapes and sizes. Behind the ear (BTE) device are fit out of the ear. Completely in canal (CIC) device, it is

completely insides the ear cannel and normally not visible. Canal (ITC) devices are fitted into the ear canal.

2.5 Factors of Hearing Loss

The following are different factors which can cause hearing loss

i. Inherited
ii. Age
iii. Use of drugs
iv. Injuries of head and ear
v. Diseases during pregnancy
vi. The level of sound (intense sound)
vii. Birth before the natural time
viii. diseases or illness

2.6 Classification of Hearing Loss

There are different types of hearing loss which suffering people are classified in the Table 2.1. Normally hearing loss is slight mild, mild, moderate, severe and profound. According to the table, slight mild people have difficulty understanding normal speech. Moderate people of hearing loss have difficulty in understanding the loud speech. Severe people can only understand the amplified speech and the profound people have difficulty understanding the amplified speech (Halawani et al., 2013).

Table 2.1: Classification of human hearing loss

Degree of hearing loss	Hearing loss range (dB HL)	Effect
Normal	-10-15db	
Slight mild	25-40dB	Difficulty understanding Normal speech
Moderate	40-70dB	Difficulty understanding Loud speech
Severe	70-90	Can understand only amplified speech
Profound	91+	Difficulty understanding amplified speech

2.7 Some Statistics about the Impaired People

The exact number of people who suffering the hearing loss is unknown in the world but there is a statistics according to (Wikipedia 2016). 71 million people in 2006 with an age of 18 to 80 in Europe were suffering from hearing loss. The Figure 2.6 showed the statistics of the Europe countries and United States who suffers from hearing loss.

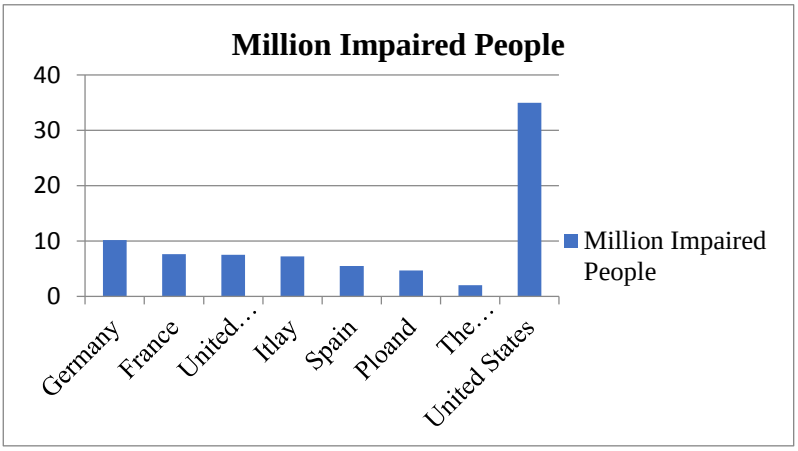


Figure 2.6: Graph of impaired people in world

2.8 Hearing Aid Devices

Hearing aids are electronic devices which receive sound wave via microphone and convert the sound wave into electrical signals; apply very complex processing during the amplification of the signals and then forward to loud speaker.

2.8.1 Analog hearing aid devices

Analog hearing aid devices only amplify all incoming sound both speech and Noise. These use small battery, microphone, speaker and simple electronic circuit which contain transistor to amplify and modify the sound which comes from outside (Halawani et al., 2013). Shown in Figure 2.7, it is cannot differentiate between the wanted (speech) signal and the unwanted (noise) signal, amplifies both sound and noise. Therefore noise can get way in a conversation. It is cannot provide any noise cancellation techniques. Some of the programmable analog hearing aid devices are also available which work better than simple (unprogrammable) analog hearing aid devices.

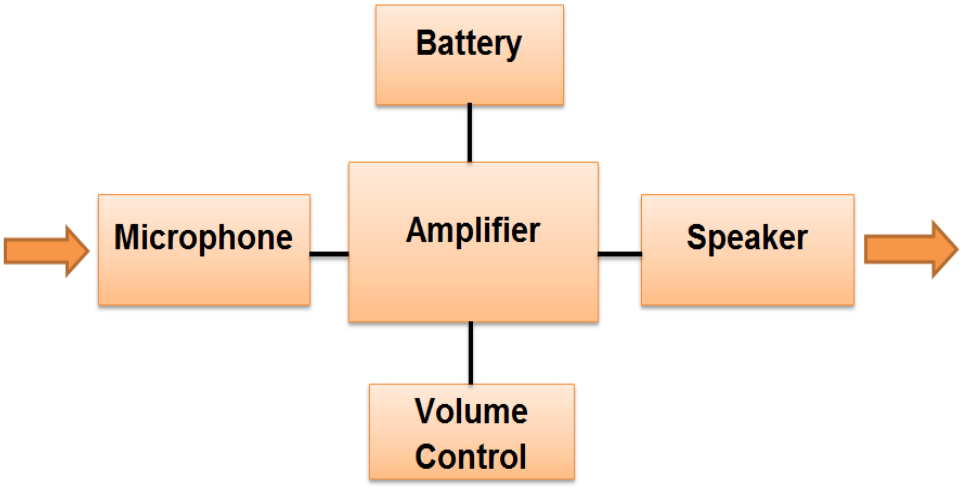


Figure 2.7: Block diagram of simple analog hearing aid device

2.8.2 Digital hearing aid devices

The digital hearing aid devices are presented for the first time to the public in 1996, which are fully digital and programmable. Digital hearing aids contain greater flexibility and adjust finely according to the patient's needs compared to the analog hearing aid devices (Halawani et al., 2013). These are used to amplify the speech signals as well as reduce the noise signals. These digital devices are also able to differentiate the speech signal and noise. Digital aids use noise reduction and speech enhancement techniques. In a digital hearing aid device, microphone receives incoming signals and converts into digital signal (1, 0) as shown in Figure 2.8. Inside the aid device, aid device use microprocessor, and small loudspeaker which forwards the income signal to the ear cannel. It is computerized chip to analyze the speech and other sound. Digital hearing aid devices are very advance using complex processing during amplification of sound and noise reduction algorithm to reduce different types of noises. Digital aids are multi memory; each memory is programmed according to their function in order to be used in specific condition and the patient can change the memory by pressing a button. Each memory works on specific place/condition; one works on normal conversation in a quiet place, other works in a noisy place and other works in a place with music or taking on mobile etc.

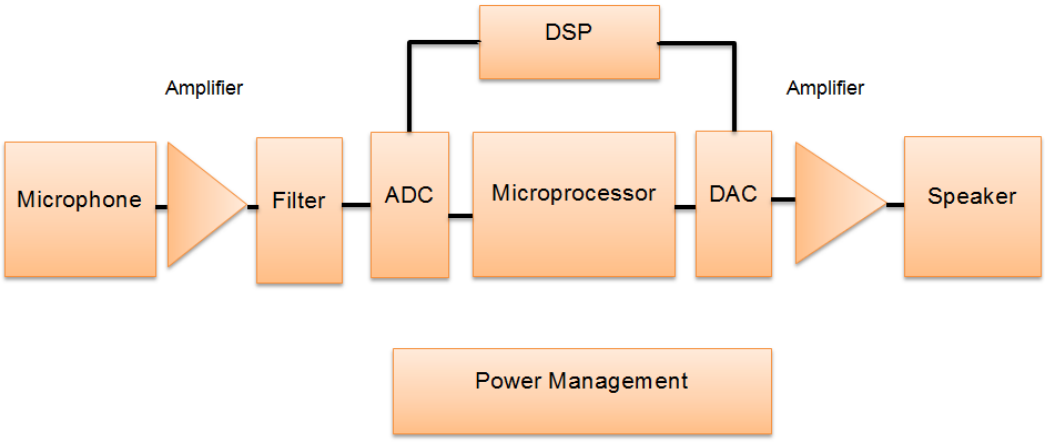


Figure 2.8: Block diagram of digital hearing aid device

2.9 Noise

Noise is the unwanted signal in speech. There are different types of noise that can affect speech signal in different way based on their time and frequency properties. Noise can be classified in the following.

2.9.1 Road traffic noise

There are people in our society who suffers from some kind of hearing loss. Vehicles on the road produce a lot of noise. Which produce a trouble for hearing impaired people. In large cities in narrow streets and high buildings, the noise is reflected which produces echo. In large cities, airways, railways, horns and whistles also increment the traffic noise and cases problems for people using hearing aids.

2.9.2 Babble noise

In our daily routine, conversations are going in the presence of multiple people. One person's speech signal is intervened with multiple talkers, which is considered as background noise or babble noise. Background talking of people is always present while talking face to face by group of people (Rodman, 2003). Babble noise always intervened with the desired speech signals which reduce the capability of the hearing of the impaired person.

2.9.3 External noise

External noises are those noises which are produced outside of the system (Rodman 2003).External noise is further divided in following types.

Atmosphere Noise: Atmosphere noise is caused by thunder storm or lighting or natural electrical disturbing occurs in nature.

Industrial Noise: Industrial noise is the noise produced due industry machine or industrial process.

2.9.4 Internal noise

Internal noise is that noise which is produced with in the system (Rodman, 2003).They are further divided into following types.

White Gaussian Noise: White noise is produced by combination of different frequencies together at once. We cannot hear white noise. In white noise, each of the frequency is projected from low to high because white noise has equal energy distribution.

Impulse Noise: Impulse noise contains unwanted instantaneous sharp sound. It is occurred by electromagnetic interference or when scratches on the recording disk or weak synchronization. It can damage human ear if it range exceed value of 180 decibels.

Impulse noise can be reduced and can improve noisy signal quality by using impulse noise filter.

Transient Noise Pulse: Transient noise is consisting of a relatively short pulse with low frequency oscillations. The first peak is caused by the impulse interference which causes resonance.

2.10 Noise Cancellation Techniques

There are two kinds of noise cancellation techniques which are Frequency Domain and Spectral Analysis. Frequency domain noise cancelation techniques look for specific types of noise frequency and remove that noise frequency through different methods. Frequency domain noise cancelation techniques are further divided into Fixed Filters and Adaptive Filters. Spectral Analysis is other technique which uses to investigate the spectrum of a signal.

2.10.1 Noise cancelation using fixed filter

Fixed Filter is used to remove a specific area of unwanted signal or noise Fixed Filters are divided into types. Each of these types has its own function and importance.

Lowpass filter passes low frequency and do not allow high frequency which increases the limit of high frequency called cut off frequency (Levitt, 2011). In Matlab we use butter filter as a lowpass filter.

Highpass filter works just the opposite of lowpass filter; it allows those frequencies which are higher than a certain limit (Levitt, 2011). It is sometimes called low-cut Filter. It is used for linear time invariant system and also use in conjunction with Lowpass Filter.

Bandpass filter is combination of highpass and lowpass filters (Levitt, 2011). It allows frequencies between two cut off frequencies. The two cut off frequency is defining at a time of filter design. In Matlab, we use Cheby 1 Filter as Bandpass Filter.

Bandstop filter is a filter that stops specific range frequencies (Levitt2011). It is opposite of the bandpass filter. Bandstop Filter is also called notch filter. In Matlab, we use ellip filter is as Bandstop Filter.

2.10.2 Noise cancellation using adaptive filter

Adaptive filter is systems which contain linear filter which have transfer function control by variable parameter its mean to adjust the using parameters according to optimization algorithm shows in Figure 2.9. Adaptive filter is very complex because of complex algorithms that is used why it as always digital. Adaptive filter is mostly used in signal processing but nowadays it is routinely used in mobile, digital camera and medical monitoring systems. Adaptive filters are used to remove noise from the speech and adjust itself according to the environment (Vanden, 1998). Adaptive filter generally use LMS algorithms. In LMS algorithms mimic a desire filter through finding filter coefficients. Those filter coefficients relate to produce LMS of the error signals .Error signals are different from the desired and the actual signals. In Matlab we use LMS algorithm for adaptive filter.

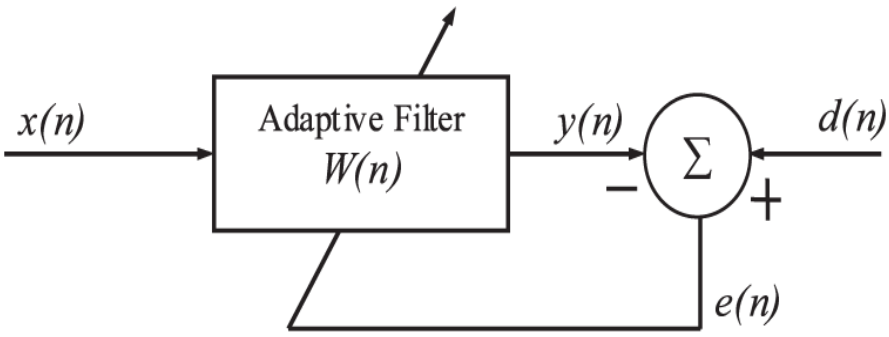


Figure 2.9: Adaptive filter

2.11 Spectral Analysis Technique

Spectral analysis means to investigate the spectrum of a signal. It has many applications in optics, speech, sonar, radar, and medicines. There spectral analysis is used for the detection of noise in speech signals (Lebart, 2001). Let the s(n) is the clean signal which is degraded by the uncorrelated noise signal v(n), then x(n) is the corrupted noisy signal which can be expressed as:

$$x(n) = s(n) + v(n) \quad (2.1)$$

Taking the Discrete Fourier Transform (DFT) of the corrupted noise signal $x(n)$:

$$x(k) = s(k) + v(k) \quad (2.2)$$

let suppose that is zero-mean and uncorrelated with, the v(n) s(n) estimate of can be |s(k)| expressed as:

$$|s(k)| = |x(k)| - e|v(k)| \quad (2.3)$$

Given |s(k)| the estimate, the speech can be expressed as:

$$|S`(k)| = |S`(k)|e^{j\theta_x(k)} \quad (2.4)$$

$$E^{j\theta_x(k)} = x(k) / |x(k)| \quad (2.5)$$

$\theta_x(k)$ is the phase of measured noisy signal. As determined by that for all practical purposes, due to computational complexity of phase of clean speech, it is sufficient to use the noisy speech phase $\theta x(k)$.

CHAPTER 3
APPLYING DIGITAL FILTERS FOR NOISE REMOVAL

As it explained in the previous section, digital hearing aid devices apply digital filters in order to remove unwanted and harmful noise from speech signals. In this section, we show the implementation results of fixed filters, adaptive filters are spectral analysis for noise removal. We will demonstrate that some of these noises cannot be removed from the speech that can be harmful for people wearing a hearing aid device. It last section of this chapter there is some samples speech that filters cannot remove the harmful noise.

3.1 Noise

Noise is the unwanted signal in speech. There are different types of noise that can affect speech signal in different way based on their time and frequency properties. Figure 3.1 shows some plots of noises types such as white noise, storm noise, high frequency noise, jet aircrafts noise and running tape.

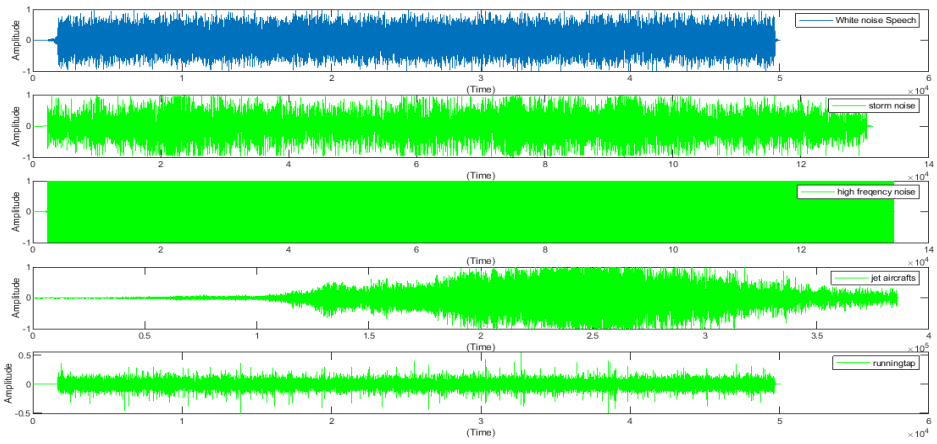


Figure 3.1: Noise types

3.2 Fixed Filter

As fixed filters, lowpass, highpass, bandpass filters can be used to remove noise from the speech signals. Each filter use different types of filter order and cutoff frequencies but in this work lowpass filter is used which can effectively remove harmful noises from the speech signal. Lowpass filters are used to remove high frequency noise from the speech signals. Lowpass filters always remove frequency near to one. The flowchart of lowpass filter shows in Figure 3.2 and algorithm are given below. For implementation, we use Matlab.

1. Record the speech signals [wavrecord]
2. Add noise to the speech signals [noise]
3. Use lowpass filter to remove noise from the speech signals
4. (n, wn)]
5. Plot the original, noise and filter signals [subplot]
6. Exit

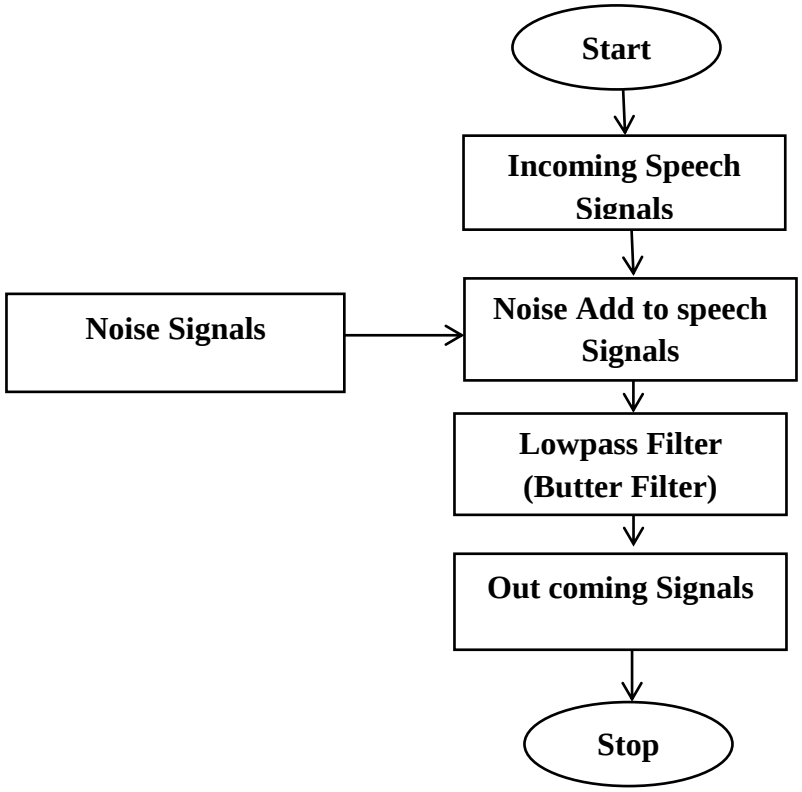


Figure 3.2: Flowchart of lowpass filter

Results of noise removal using lowpass filter is shown in Figures 3.3 and 3.4. As it can be seen, results are different from one another because of different types of noises. In both, 6^{th} order filter with cutoff frequency of 0.6 is used with butter function as lowpass filter. In Figure 3.3, first we plot the speech signal. Secondly jet aircraft noise is added to speech signal, and then lowpass butter filter is applied to remove jet aircraft noise from the speech signal. While in Figure 3.4, first we plot the speech signal. Secondly some high frequency noise is added to speech signal, and then lowpass butter filter is applied to remove fixed high frequency noise from the speech signal. lowpass filters are used to remove high frequency noise from the speech signals. Finally plotting in Figure 3.3 and 3.4 shows the result of the signals after noise removal.

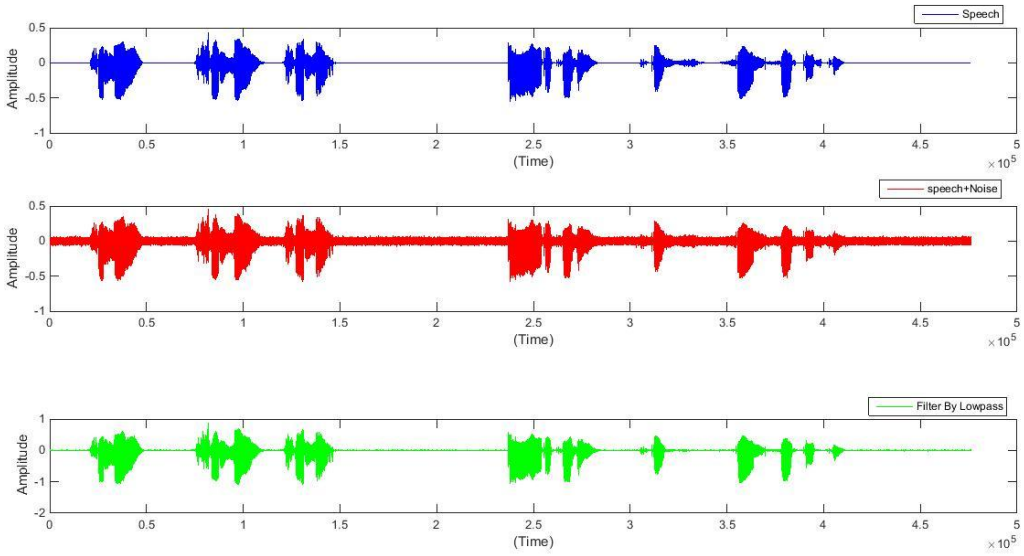


Figure 3.3: Gaussian noise cancellation using 6^{th} order lowpass filter with 0.6 cutoff frequencies

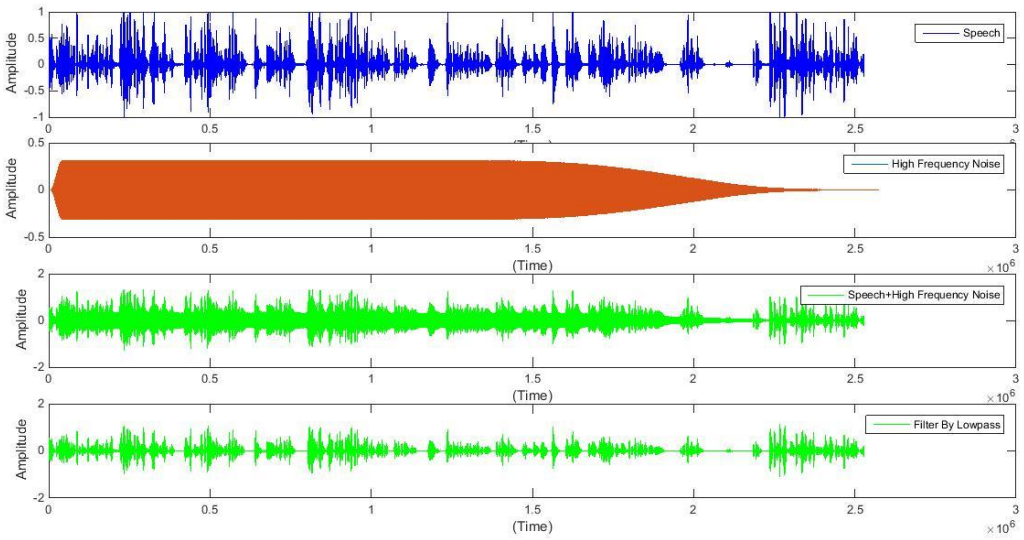


Figure 3.4: High frequency noise cancellation using 6th order lowpass filter with 0. 6 cutoff frequencies

3.3 Adaptive Filter

Adaptive filter is systems which contain linear filter which have transfer function control by variable parameter its mean to adjust the using parameters according to optimization algorithm. Adaptive filter is very complex because of complex algorithm that is why it is always digital. Least mean squares (LMS) algorithms are a class of adaptive filter used to mimic a desired filter by finding the filter coefficients that relate to producing the least mean square of the error signal (difference between the desired and the actual signal).The flow chart show in Figure 3.5 and algorithm are given below.

1. Record the speech signals [wavrecord]
2. Add noise to the speech signals [awgn]
3. Use LMS adaptive filter to remove noise from the speech signals [adaptfilt.lms]
4. Plot the original, noise and filter signals [subplot]
5. Exit

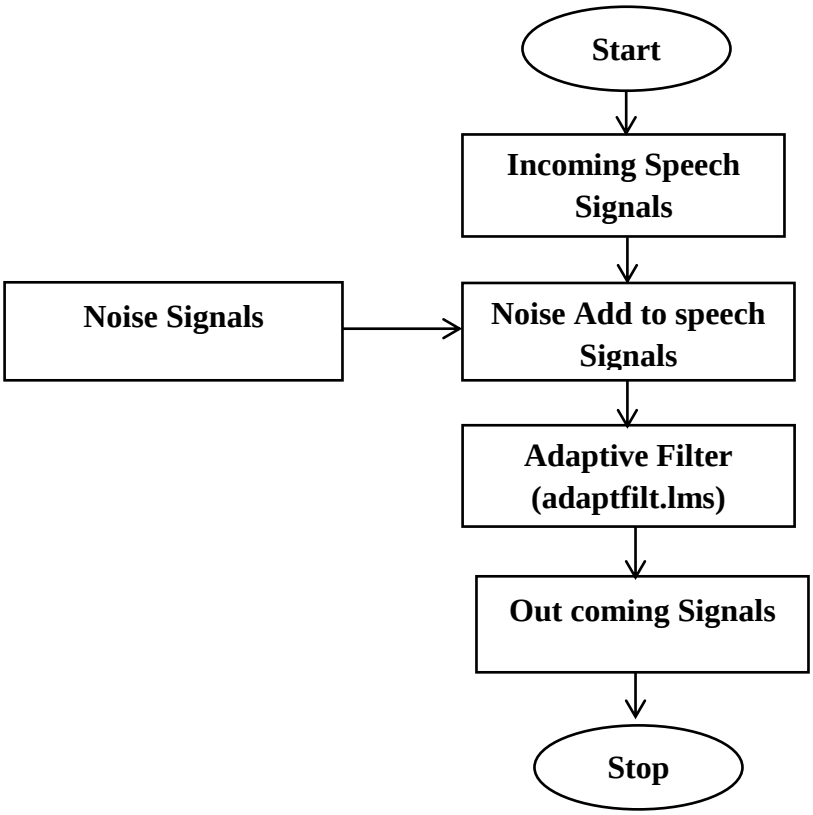


Figure 3.5: Flowchart of adaptive filter in noise cancellation

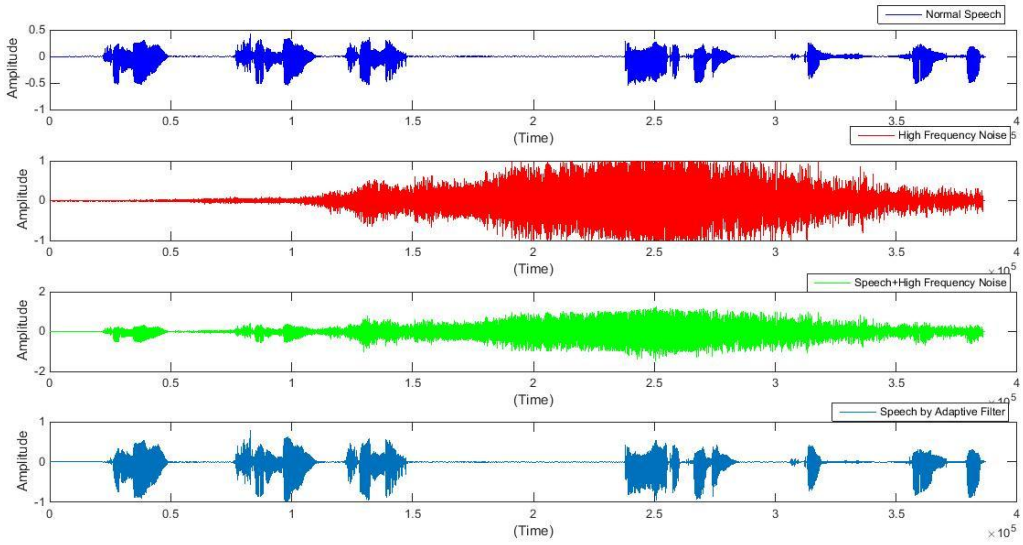


Figure 3.6: Jet aircrafts noise cancellation using adaptive filter with step size 0.001

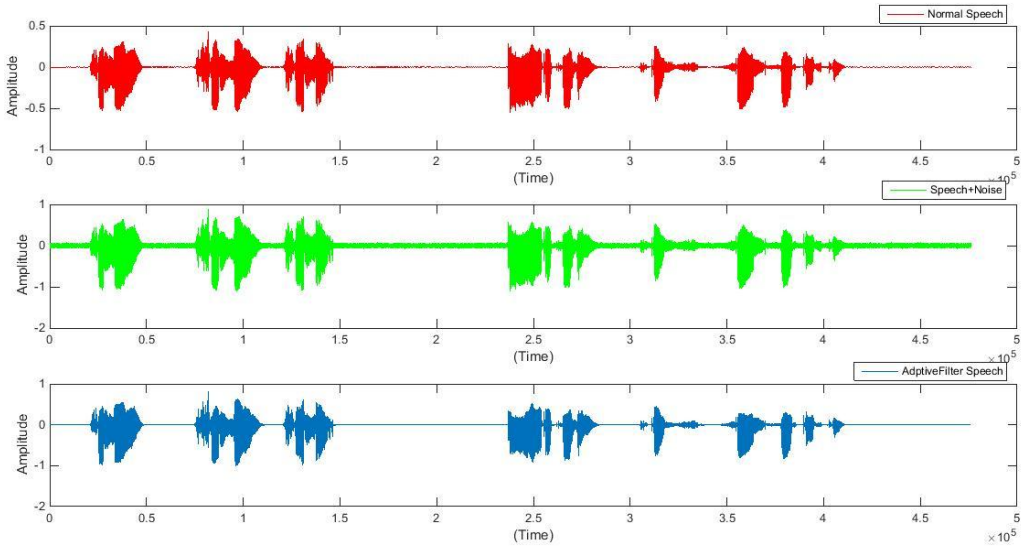


Figure 3.7: Gaussian noise cancellation using adaptive filter with step size0.11

Results of noise removal using adaptive filter is shown in Figures 3.6 and 3.7. As it can be seen, results are different from one another because of different step size in LMS algorithm. In Figure 3.6 shows the step size is0.001 of the LMS algorithm while in Figure 3.7 the step size is 0.11 of the LMS algorithm. In both Figures, LMS adaptive filter is used. In Figure 3.7 first we plot the speech signal. Secondly Jet aircrafts Noise is added to the speech signal, and then used adaptive filter to remove Jet aircrafts noise from the speech signal. In Figure 3.6 first we plot the speech signal. Secondly some fixed high frequency noise is added to the speech signal, and then used adaptive filter to remove fixed high frequency Noise from the speech signal. Background noise change by time to time the adaptive Filter is able to adjust itself according to the background noise. Finally plotting the all result of the signals.

3.4 Spectral Analysis

Spectral analysis means the analysis of the spectrum of a signal with respect to frequency. There spectral analysis used to remove a specific frequency. The algorithm of spectral analysis is given below.

1. Record the speech Signals [wavrecord]
2. Use [FFT] function
3. Add noise to the Speech signals [awgn]
4. Plot the original, Noise and filter signals [Subplot]
5. Exit

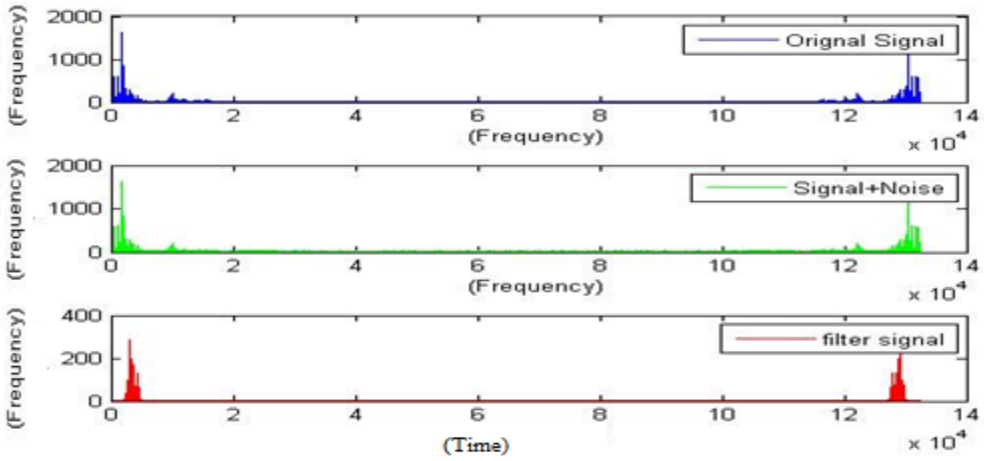


Figure 3.8: Noise cancellation using spectral analysis

Spectral Analysis use FFT function. In the techniques we move from time domain to frequency domain. It is shows in Figure 3.8 first fist plot the speech signal then add noise to the same signal. And in last filter by lowpass filter.

3.5 Comments

Here some samples of the speech signals which is corrupted by the noise. It is clear when we apply lowpass and adaptive filter techniques to remove noise as result we recover the original speech signal but we also lost some data of the original speech which shows in Figure 3.9 and 3.10. It is the disadvantages of these filters. That's why we move toward the CNN networks to classify data.

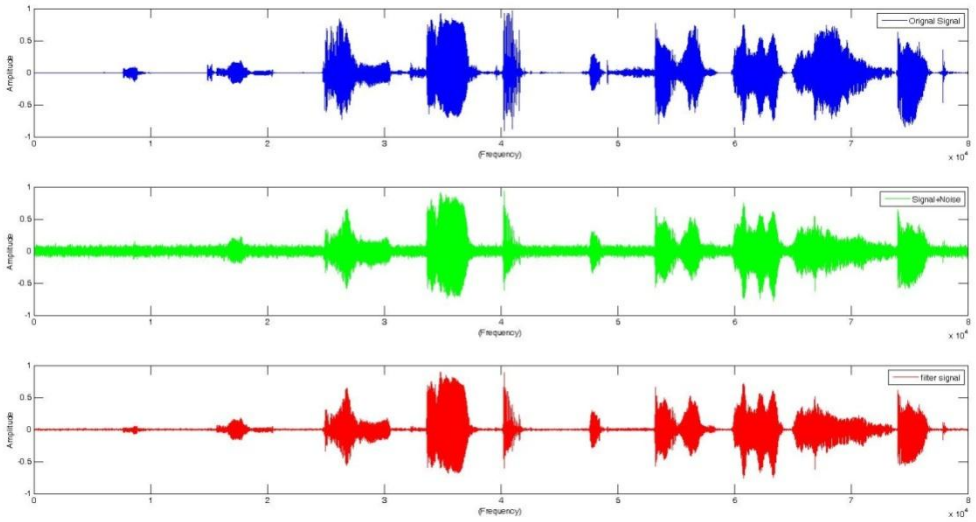


Figure 4.9: Gaussian noise cancellation using lowpass filter

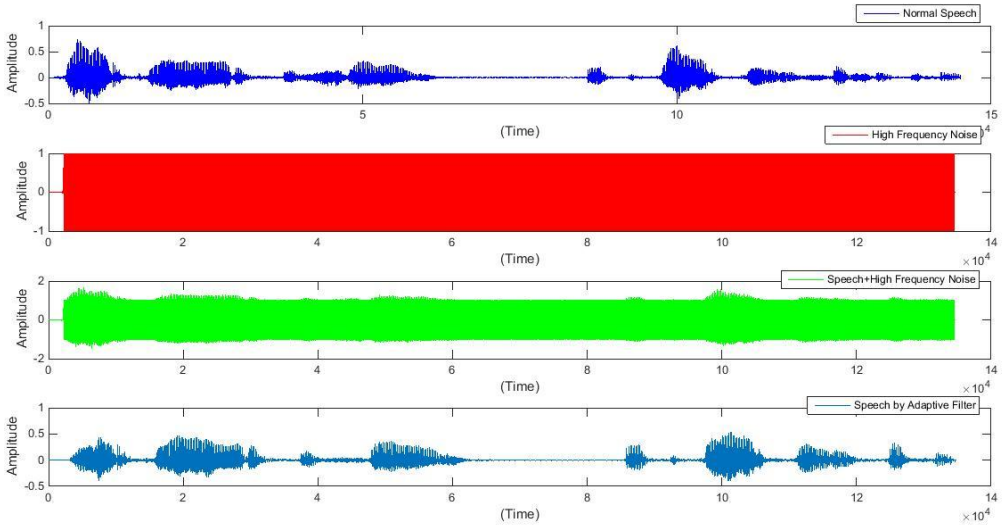


Figure 4.10: High frequency noise cancellation using adaptive filter

CHAPTER 4

SPEECH SIGNALS AND CONVOLUTIONAL NEURAL NETWORKS

In this chapter, first speech signals and how these signals are converted to spectrogram images are explained. Then, basic layers of Convolutional Neural Networks (CNN) are discussed.

4.1 Speech Signals

Speech is a way of communication to convey information from one human to another. There are some properties of speech signal. Speech signal is one dimensional in nature and independent with respect to time variable. Speech signals is non-stationary in nature which means that frequency domain of the signals does not remain constant with respect to time variable.

The sound range which human ear can hear is from 20 Hz to 20 KHz .The range of speech signal is 1000 Hz to 5000 Hz and the bandwidth of speech signal is 4 KHz as shown in Figure 4.1. But for the speech signal the sensitivity of human ear is at its peak when the speech signal range is between 400 Hz to 3400 Hz (Halawani, al, 2013).

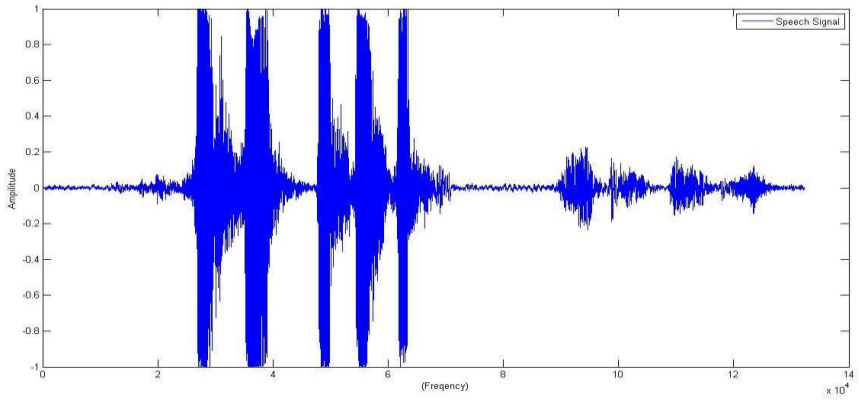


Figure 4.1: Speech Signal

4.1.1 Conversion of Speech Signals to Spectrogram Images

In order to categorize the speech signals using convolutional neural networks, first the speech signals are converted to two dimensional images using spectrogram. A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. In particular, to analyse the spectral audio, Fourier transform is is very useful (Ramírez, 2107).Signals are converted from the time domain to the frequency domain using Fourier transform in order to obtain the magnitude and phase of each frequency. Fast Fourier Transform (FFT) is an optimized implementation of the Discrete Fourier Transform (DFT). This can effectively be used for real-time analysis on discrete sequences, such as sample values in audio. In short, FFT correlates frequencies contained in the signal and bins them together in discrete steps. Particularly a sliding FFT window is uses by STFT to obtain spectra for each segment in time of the original signal. The squared magnitude of each spectra which is obtained is then stacked together to form a power spectrogram estimate.

The Discrete Fourier Transform (DFT) defines as:

$$S(k) \quad = \sum_{n=0}^{N-1} s(n) e^{-j\frac{2\pi}{N}kn} \tag{4.1}$$

Here *k*refers to the frequency bin or Fourier component number, and *nis the* sample number.In general, $S(k)$ is complex valued, and usually the only interesting part is the magnitude of it. It is refers a magnitude spectrum $S(\omega)$. $S(\omega)^2$ is the power spectrum. Which is obtained to calculated squaring the magnitude spectrum..

FFT is used effectively for the real time analysis of the discrete sequences. To reveal the information about the audio Spectrogram is used (Lin et al., 2013). The information that can be obtained by using the spectrogram is about the harmonics, time variant events, periodic events and temporal localization of the events which show clearly in Figure 4.2. Location of events in time and frequencycontent are clearly visible, compared to Figure4.1.

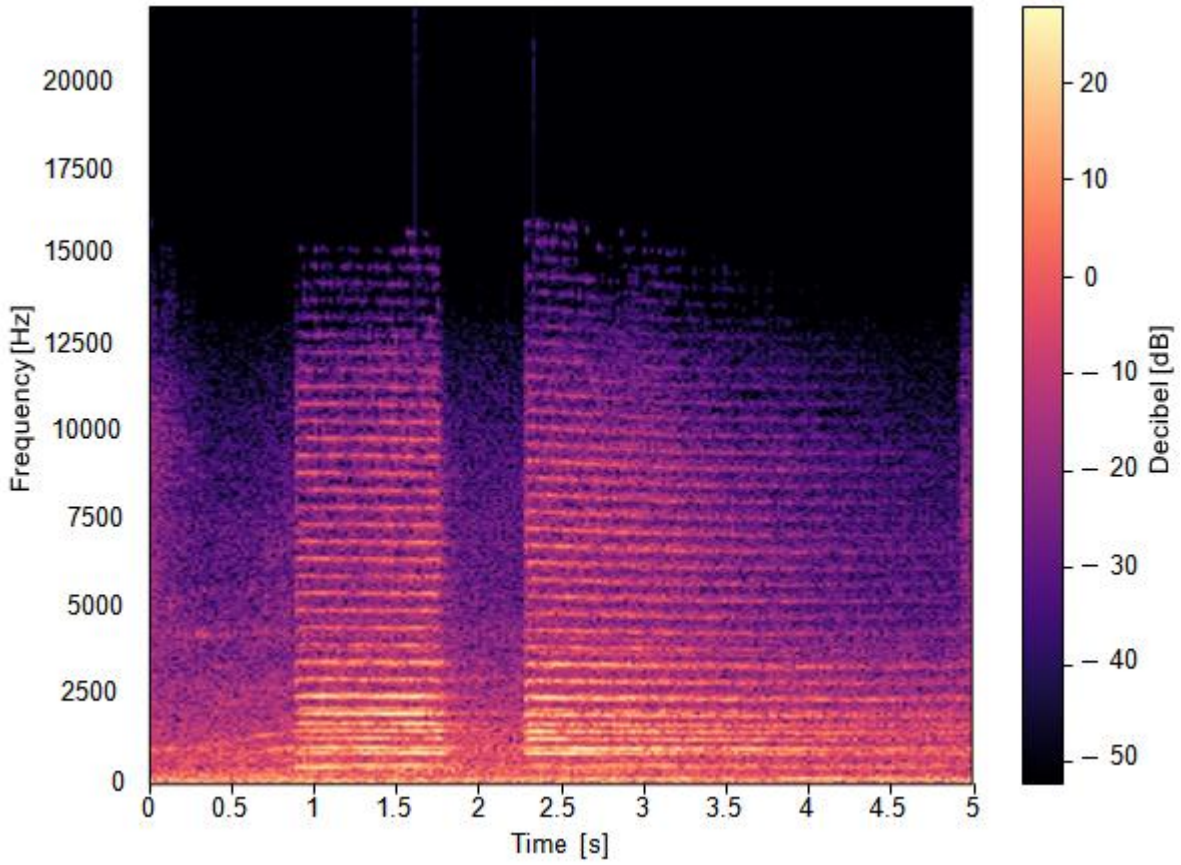


Figure 4.2: Spectrogram image of a car horn.

4.2 Convolutional Neural Networks

In the field of deep learning, a convolutional neural network (CNN) is a class of deep neural networks, which is mostly applied to analyzing visual imagery. convolutional neural network is a regularized form of multilayer perceptrons which is usually refer to fully connected networks. In fully connected network each neuron in one layer is connected to all neurons in the next layer. (wekipedia 2019). The sample architecture of CNN is show in Figure 4.3 which has input images, convolution layer, pooling, fully connected layer and output

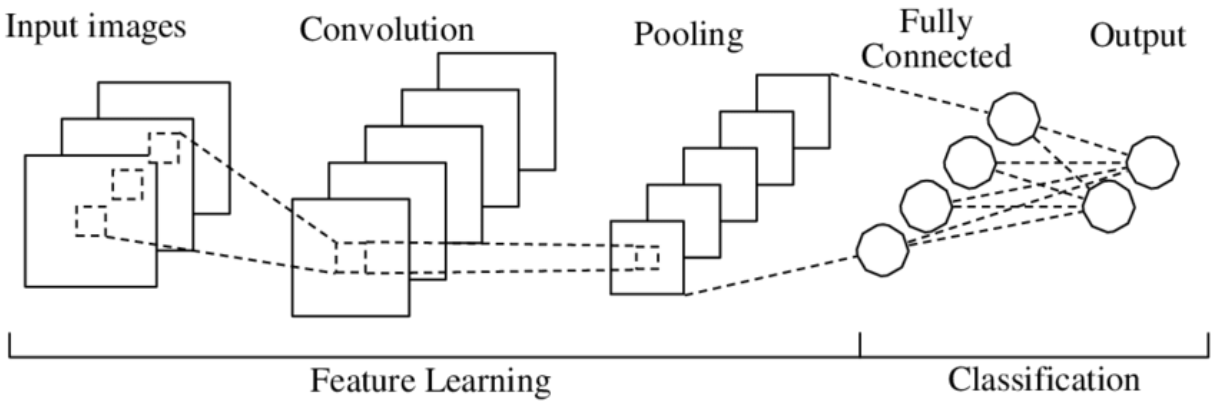


Figure 4.3: General framework of CCN

.

There is learning alogrithms categorized betweem supervised versuc and unsupervised algotithms. Suprvised alogrithms are work with with labled data and unsupervised alogrithms work with unlable data.

Supervised algorithms are provided with desird output (calss lables), in order to continuing the learing process and reduce the classification error as well as to updata the model.In summary , a supervise classification alogrithm can be formally expressed:

$$f(x,w)=\hat{y} \qquad (4.2)$$

where x is the input to be classified , w are the learned parametes of the fuction, and $\hat{y}$ is the predicted class lable

4.2.1 Convolutional Layer

Convolutional Neural networks works on layer to layer operation called convolution. Convolution is a mathematical operation, just like addition, multiplication and integration. In multiplication we take two number and produce a third number, same like multiplication in convolution operation convolution takes two signals and produce a third signal.

The following equation shows the discrete convolution in 2-D of two images, I and K as:

$$F(i,j) = (I * K)(i,j) \quad = \sum_m \quad \sum_n I(\mathrm{m},\mathrm{n})\mathrm{k}(\mathrm{i}-\mathrm{m},\mathrm{j}-\mathrm{n}) \tag{4.3}$$

Here *m* and *n* are the dimensions of the filter (kernel). *i* and *j* are represented the indexes for each pixel. In convolutional network, *I* is the first argument to the convolution is often referred to as the input and *K* is the second argument referred to as the kernel (filter). *F*is feature map referred to as the output.

$$F(i,j) = (I * K)(i,j) \quad = \sum_m \quad \sum_n I(\mathrm{m},\mathrm{n})\mathrm{k}(\mathrm{i}+\mathrm{m},\mathrm{j}+\mathrm{n}) \tag{4.4}$$

Generally,Cross-correlation is the operation applied in convolutional layers, and not convolution. Figure 4.3 shows the example.*I* is input is convolved with a kernel (*K)*and a feature map*(I*K)* is obtained. The number of feature maps*(F)* will depend on the number of kernels (filters) that the designer selectsfor designing of each particular convolution layer. In convolutional network the last layer is regular fully connected layers. 2-Dcross-correlationis shows in Figure 4.3 betweenaninput *I*andakernel*K*, producing a feature map*K*. The red area which highlighted is refer to the receptive field ofthe input.

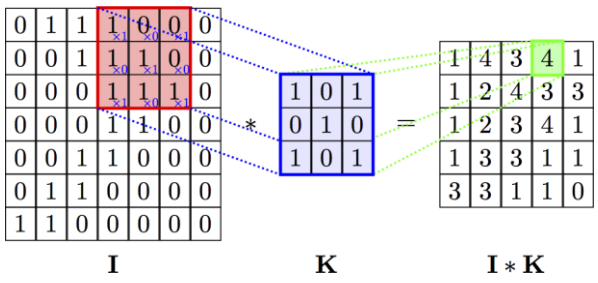


Figure4.4:2-Dcross-correlation of CNN

4.2.2 Pooling layer

To produce a nonlinear representation, a typical CNN use the activation function. The pooling layer is used to modify the output further. Actually the spatial size of the representation is reduced by pooling layers; reduce the computing and the number of parameters. On the neighbourhood values, the output of the network is replaced by a statistical operation at a certain position. The pooling layer applies several operations to the neighbourhood of a

location. The operation applied by the Pooling layer can be the weighted average, one square norm and the maximum average. Among them the most popular operation is the MAX Pooling due to its practical application. It also returns the maximum value within neighbourhood of the values. The following figure shows an example of the MAX Pooling in which the action reduces the dimension of the input data (Dumoulin et al., 2016).

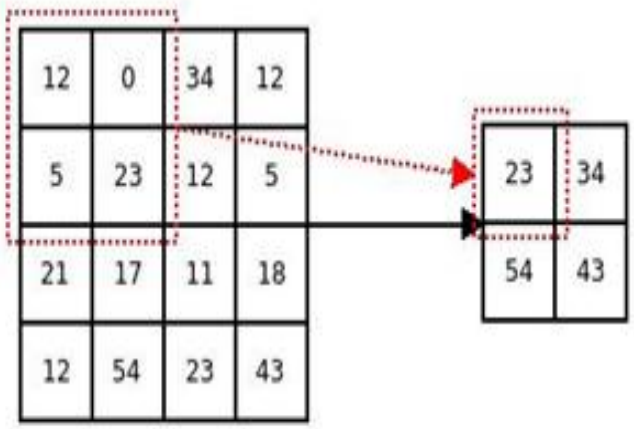


Figure 4.5: Max Pooling

4.2.3 The fully connected layer

The fully connected layers are the last layers in a convolution network. The capability to learn the nonlinear combination of the features learned earlier in the network is actually provided by this fully connected layer. Interaction is done by every parameter in this network with its own part of the input. To compute the final output a term is added after the successful employment of a term in the convolution layer. In Figure 4.6 the illustration is structured for the fully connected layer. In the structure of the fully connected layer the edges show the learned parameters which are seen to be multiplied with the input (Wikipedia 2018).

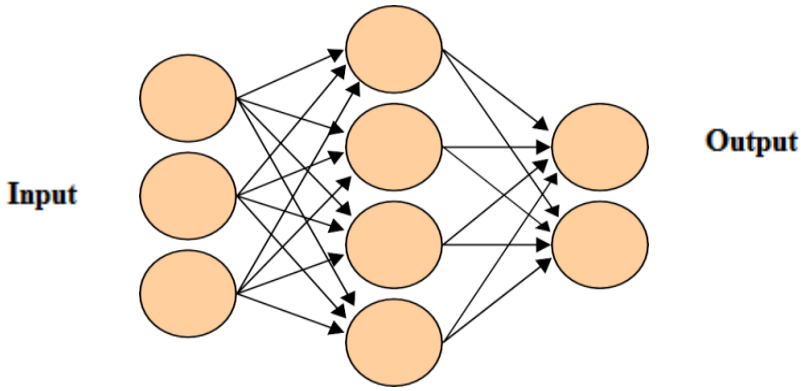


Figure 4.6: Illustration of the fully connected structure

4.2.4 The Rectified linear unit

The Rectified linear unit is currently the most popular activation function at the time. For the modelling of the complex relationships, the several linear transformations are not enough on top of each other. As we know that convolution and matrix multiplication in fully connected layer are linear operations we required the addition of some non-linearity's to our models. We can avoid the collapsing of the model by applying an activation function to each output layer. The Rectified linear units have some advantages such as efficient computation and efficient gradient propagation (Ramachandranet al., 2017) these properties make the Rectified linear units to be used as the most popular activation functions among other activation functions in the line.

A ReLU layer performs a threshold operation to each element of the input, where any value less than zero is set to zero.

This operation is equivalent to

$$f = \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases} \quad (4.5)$$

4.2.5 Soft max activation

Soft max activation is the generalization of the logistic function to multiple classes. Therefore, the output represents a categorical distribution over multiple classes. The soft max function takes a vector **x** of arbitrary values and squashes them to a vector in the range [0,1] that sums to 1.

The formula for soft max is:

$$\sigma(x)j = \frac{e^{xj}}{\sum_{k=1}^{K} e^{xk}} \tag{4.6}$$

for j = 1, . . . ,K. Where K is the number of classes.

The values can now represent a probability distribution over *K* different outcomes.

4.2.6 Parameter initialization

A batch normalization layer normalizes each input channel across a mini-batch. Batch normalization layers are used between convolutional layers and nonlinearities to speed up training of convolutional neural networks. It also reduces the sensitivity to networks initialization, such as ReLU layers.

All trainable parameters of CNN layers need to be initialized in some manner. The trained parameters are initialized in a manner that bias variables are zero.

$$W_{ij} \sim U\left[\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right] \tag{4.7}$$

where *n* is the number of columns/rows in *W* and *U* is the uniform distribution.

4.2.7 Optimization

If we are able to define the error of each predicted class during the training, we shall be able to find the optimal parameter setting for a given model. This is achieved by defining a loss function. This function is minimized when the parameters of this model are adjusted. And in this way the optimal parameter can be obtained. We are using the Cross entropy loss among many loss functions. This type of loss is chosen because it is the most common loss for classification tasks.

4.2.8 Regularization

A machine system is said to be the most successful system if it achieves low training error, and minimize the difference between test error and training of the system. Training error is the measure of error over that particular data that the system uses for training. The unseen data can help derive the test error. Test error is used to find the performance of the system. For some task when the model is complex, it is that condition when the training error of the model is low and the test error is high. This type of situation is termed as over fitting.

4.3 Speech processing and Related Work

According to (Palaz, et al., 2015) the speech recognition and the neural predictive coding using the convolutional coding. They proposed an unsupervised technique to learn speaker specific characteristics from unlabeled data. Two sets of evaluation experiments were performed (closed set speaker identification and a large scale speaker identification experiment). During their experiments the NPC embedding outperformed the i-vector at the frame level speaker identification. At the utterance level speaker identification task, this also provided complementary information to the i-vectors. Generally the resulting embedding captures significant information about the speaker's identity.

According to (Jati, et al., 2018) analyzed the Raw speech as input using the CNN-based speech recognition system. They studied to evaluate the susceptibility of the CNN-based system to mismatch the conditions. They investigated the aspect into TIMIT phoneme

recognition task and Aurora2 connected work recognition task. Their work suggested that to improve the robustness of the CNN-based system, the parallel between the time domain processing and the frequency domain processing can be exploited. They also presented a method to understand the speech information by computing the mean frequency responses of the filters

To studied and discussed the Speech command recognition with the convolutional neural network used three models (Vanill, DNN and CNN) of which the convolutional neural network outperforms the others with more accuracy According to (Li, X, et al). They intended to study the new CNN architectures with few multiples to work on power-constrained devices.

According to (Chang, et al., 2014) to study the Robust CNN-based speech recognition with the Gabor filter kernels. In their comprehensive study, they proposed a neural architecture called "Gabor Convolutional Neural Network Architecture". There experiments used two noisy versions of the WSJ corpus (Aurora 4 and RATS re-noised WSJ). In both of the noisy versions the GCNN with Gabor features performed better than ETSI-AFE, PNCC and the Gabor without the CNN approach.

CHAPTER 5

CLASSIFICATION OF HARMFUL SPEECH SIGNALS USING CONVOLUTIONAL NEURAL NETWOKS

5.1 System Architecture

Convolutional Neural Network performs operations on images using multiple layers. Figure 4.1 shows the block diagram how CNN classifies speech spectrogram images using different layers. Input to the network is speech spectrogram images and speech+noise spectrogram images. These input spectogram images (28x28x1) Pixel. In the first layer of convolution, apply the convolution operation with 8 filters of 3 x 3 with ReUl layer, output will become 26 x 26 x 8. So Then apply pooling with 2 x 2 filter to minimize the size to 13 x 13x 8 of the image. In the convolution second layer, apply the convolution operation with 16 filters of size 3 x 3. The output become 11 x 11 x 11 with ReUl layer then on which apply pooling layer with 2 x 2 filters and the size l minimize to 5 x 5 x 16. Finally, the input size 5x5x16 size of image passes it through fully connected layers and output size is the number of classes which is 2. Fully connected layer convert our image matrix into a classification matrix to obtain the output result.We have tested different filter sizes as shown in Evaluations chapter.

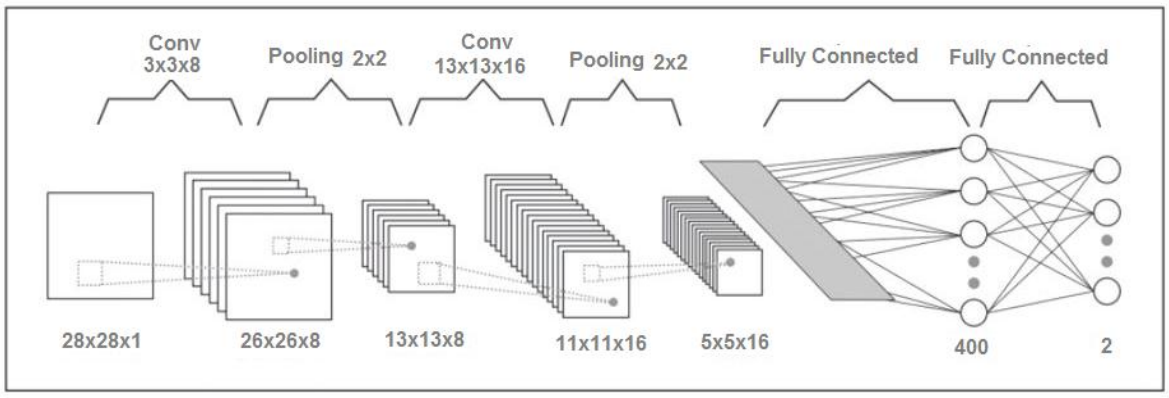
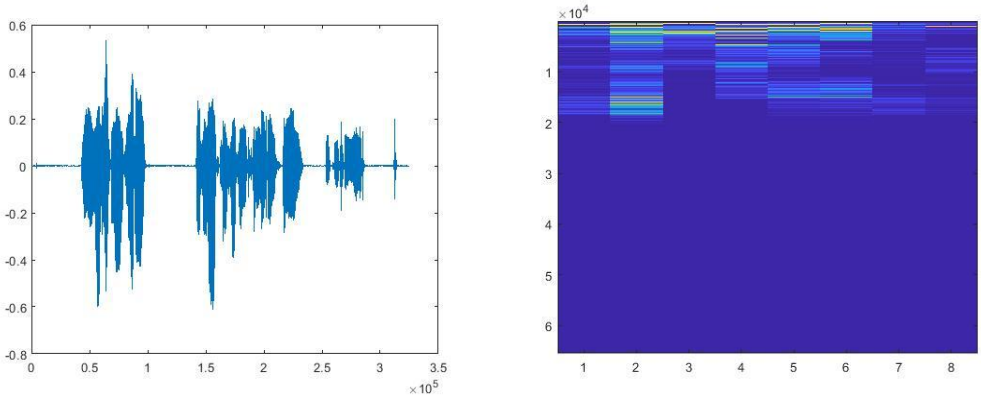


Figure 5.1: Block diagram of the CNN classification of harmful noise

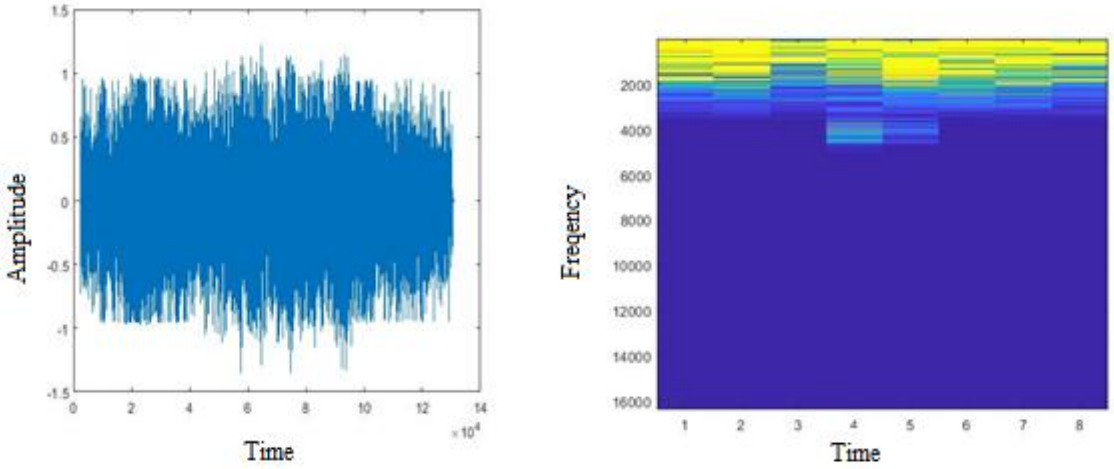
5.2 Speech and Noise Signal Spectrograms

Speech signals are one dimensional. Generally, speech signals are converted to two dimensional images in order to classify with CNN. That is why we convert the speech and speech+noise signal to spectrogram images using spectrogram function. Spectrogram of the speech signal is shown in Figure 4.2. Here we add jet aircrafts noise to the speech signal and the spectrogram of noise signal is shown in Figure 4.3. For the same speech signal, spectrogram of with the added high frequency fixed noise is shown in Figure 4.4. As can be seen, spectrogram images are changing according to the added noise type. These spectrogram images are labelled as clean speech and noise speech and are given to the CNN network for classification. For each noise type, we train a separate CNN network with the clean speech spectrogram image and speech+noise spectrogram image for classification. We can use both gray scale and RGB images for CNN model. But here in our work we use gray scale imges insead of RGB to train the CNN. It is very important to note that the model to work on gray-scale or colored images in the future.



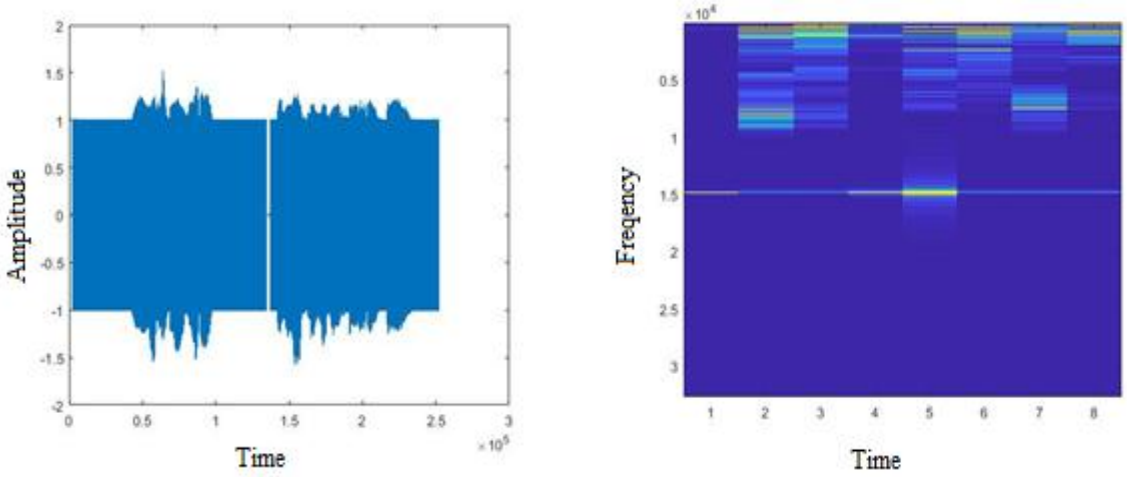
(a) Speech Signal (b) Spectrogram image

Figure 5.2: spectrogram of the speech signal



(a) Speech + jet aircrafts noise (b) Spectrogram image

Figure 5.3: spectrogram of the jet aircrafts noise added to speech signal



(a) Speech + High frequency fixed noise (b) Spectrogram image

Figure 5.4: spectrogram of the High frequency fixed noise added to speech signal

5.3 Input Spectrogram Images to Convolutional Neural Network

We load spectrogram gray scale image data as an image-data-store. Image-data-store loads stores and automatically label them based on folder names. It is very important to note that during the training of CNN network, the image-data-store efficiently reads batches of images.

Another important factor is specification of the size of the images in the input layer of the network. For current work, the input image size is 28-by-28 pixels. Figure 5.5 shows the size of spectrogram gray scale image of the speech and white noise by 28-by-28 pixels. Figure 5.6 shows spectrogram gray scale image of the speech and high frequency noise (28-by-28). Finally, Figure 5.7 shows spectrogram gray scale image of the speech and storm noise (28-by-28).

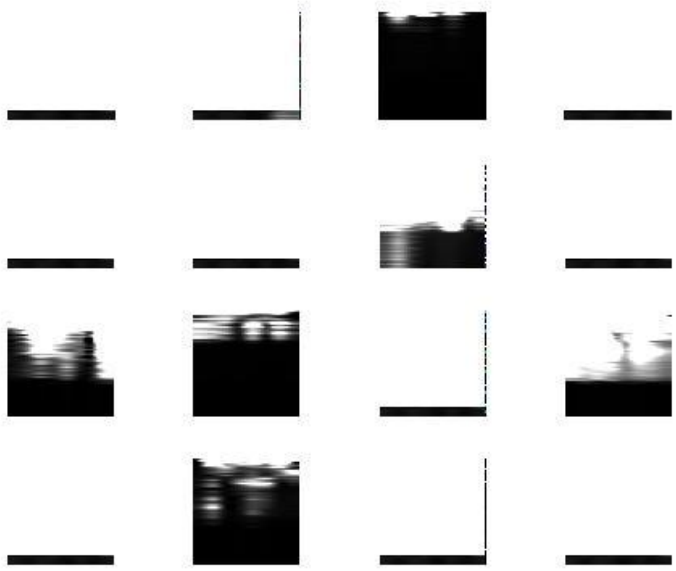


Figure 5.5: spectrogram gray scale image of the speech and white noise

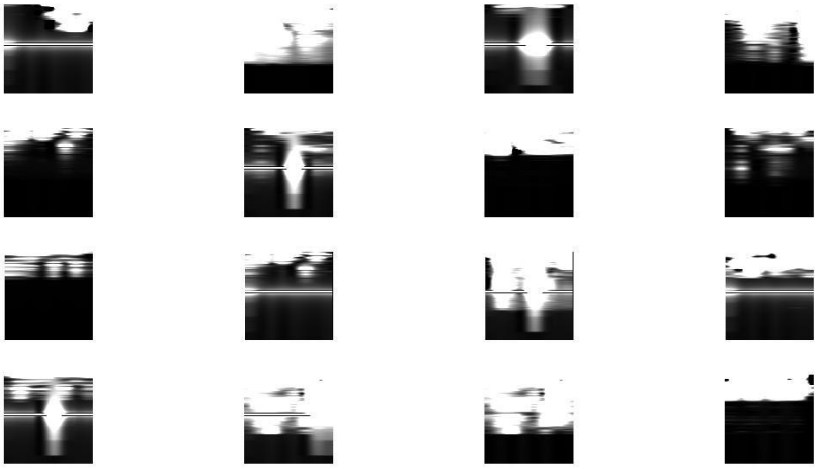


Figure 5.6: spectrogram gray scale images of the speech and High frequency noise

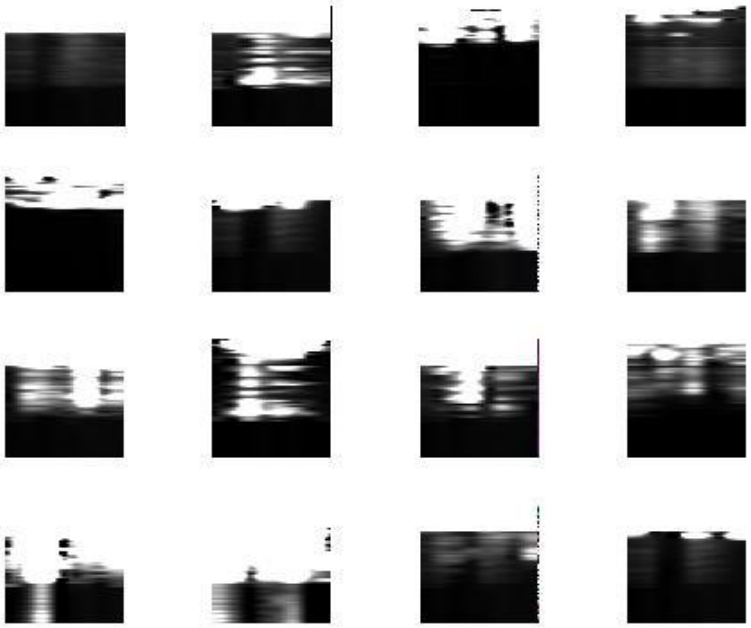


Figure 5.7: spectrogram gray scale images of the speech and storm noise

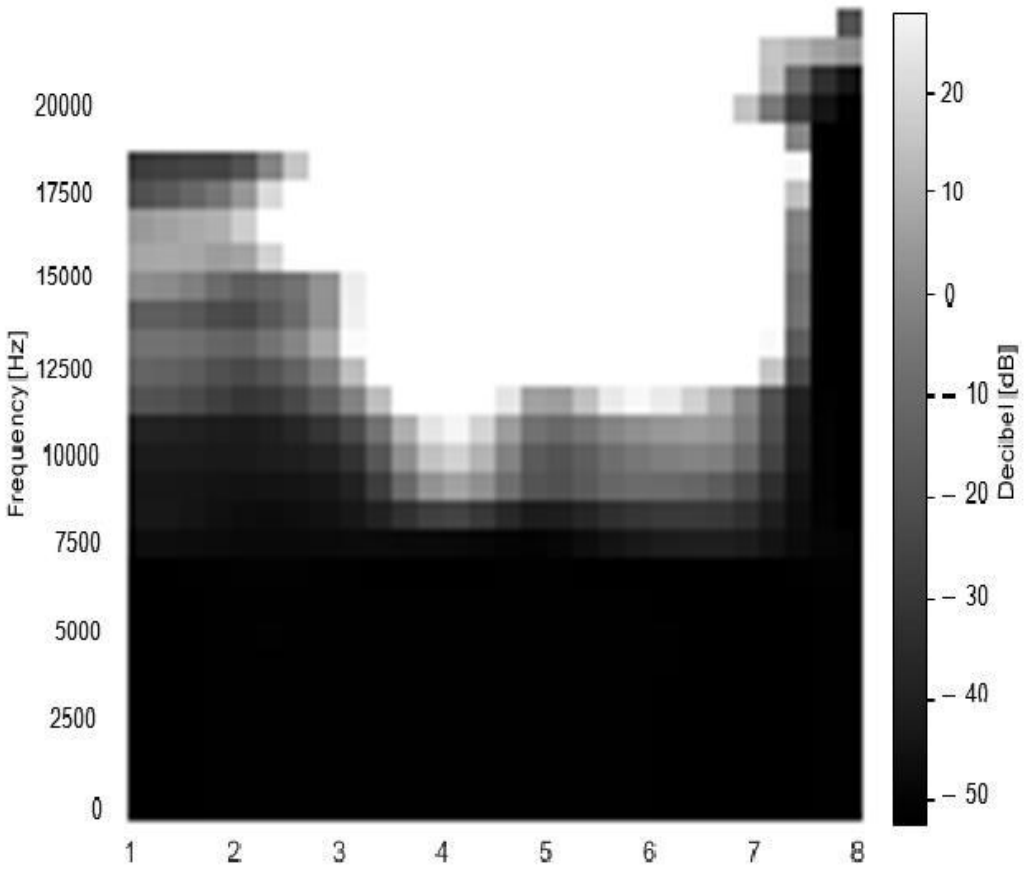


Figure 5.8: 28 by 28 pixel input image to CNNs network

5.4 Defining Convolutional Neural Network Architecture

The CNN architecture for harmful noise classification is given in Figure 5.8. An imageInputLayer here specify the gray scale image size is 28 by 28 pixel and channel color is 1. For colour image convolutional neural networks use the channel size of 3 and for gray scale image the channel size is 1. First augment is Filter size is 7-by-7, where we use 7 by 7 filters. The second argument is numFilters (number of filters), it mean the number of neurons. Theses neurons connect to the same region of the input. The number of feature maps determined by this parameter. We tested different number of neurons as shown in the next subsection. To add padding to the input feature map, CNN uses padding value name pair

function. By default stride of 1, same padding for convolutional layers which means the output spatial size is the same as the input.

ReLU is the most common activation function which has the complete name as rectified linear unit. A ReLU layer performs a threshold operation to each element of the input, where any value less than zero is set to zero. In our CNN architecture, ReLU layer is used which is followed by the batch normalization layer. In max pooling layer, first pool size is specified by the first argument, which returns the maximum values (regions rectangular) of the inputs. In current work rectangular region size is used by 2x2 filters. The step size is specified by the stride name value pair arguments; the step size takes training function and scans along the input. One or more than one fully connected layers can be used in CNN networks. Therefore the convolutional layer followed by the fully connected layers. In our work, we use one fully connected layer. The parameters of the outsize in the last fully connected layer must be the same as the number of classes in our target data. In this work the output size is 2. This is corresponding to the 2 classes of clean speech and harmful noise.

5.5 Train Network Using Training Data to Classify Validation Images and Compute Accuracy of Speech and High Frequency Fixed Noise

The following graphs show the training progress, the mini batch loss, accuracy, validation loss and accuracy for the high frequency noise.

5.5.1 CNN using One Convolutional Layer

The block diagram of the CNN network architecture is shown in Figure 5.8; in particular one convolution2dLayer, one maxPooling2dLaye and one fully connected layer is used. The Figure 5.9 is the training progress plot. This plot has two parts one for accuracy and other for loss, which contains the validation loss, accuracy/mini-batch loss and accuracyfor the clean speech and speech+noise signals. Here we use convolution2dLayer with filter size of 7x7 and channel size is 8 with the same padding, 2x2 maxPooling2dLayer with (stride2) and fullyConnectedLayer. After training the network the classification accuracy is 99.88%.

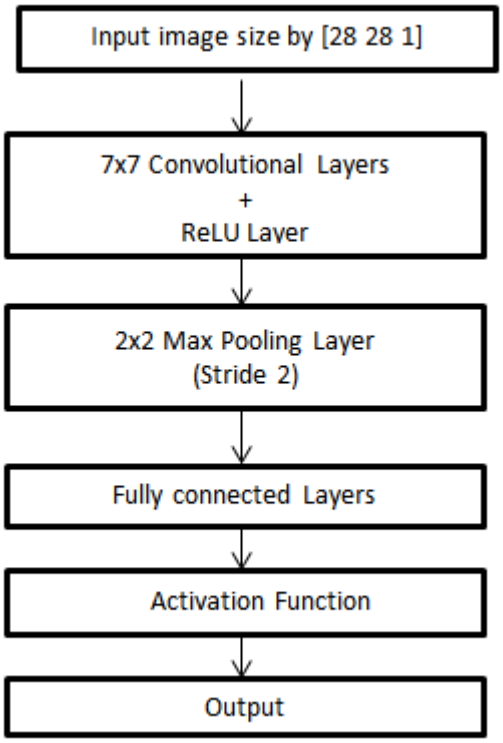


Figure 5.9: Block diagram of CNN for classification of harmful noises using one layer

.

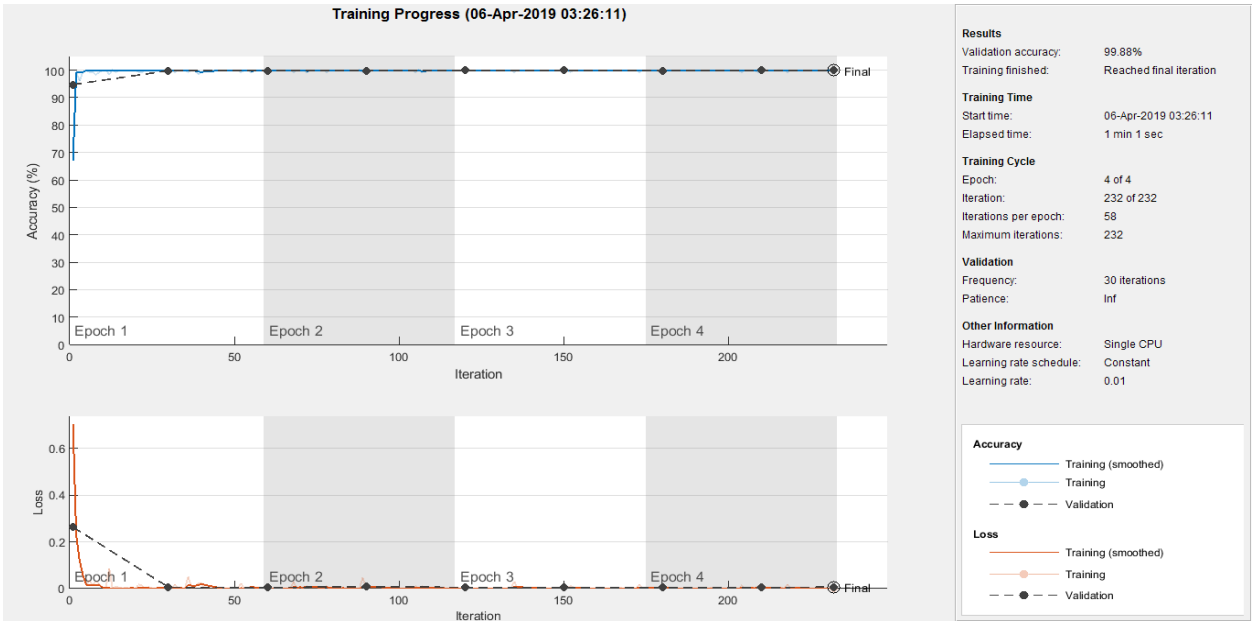


Figure 5.10: Validation accuracy of the speech and high frequency noise using one layer

5.5.2 CNN using Two Convolutional Layers

The block diagram of the CNN network architecture is shown in Figure 5.10; in particular two convolution2dLayers, two maxPooling2dLayers and one fully connected layer is used. The Figure 5.11 is the training progress plot. This plot has two parts; one for accuracy and other for loss. Here we use first convolution2dLayer with filter size is 7x7 and channel size is 8 with the same padding, 2x2 maxPooling2dLayer with (stride2) .Then we apply second convolution2dLayer with filter size is 5x5 and channel size is 16 with the same padding, 2x2 maxPooling2dLayer with (stride2) and fullyConnectedLayer. After training the network the classification accuracy is 99.96%.

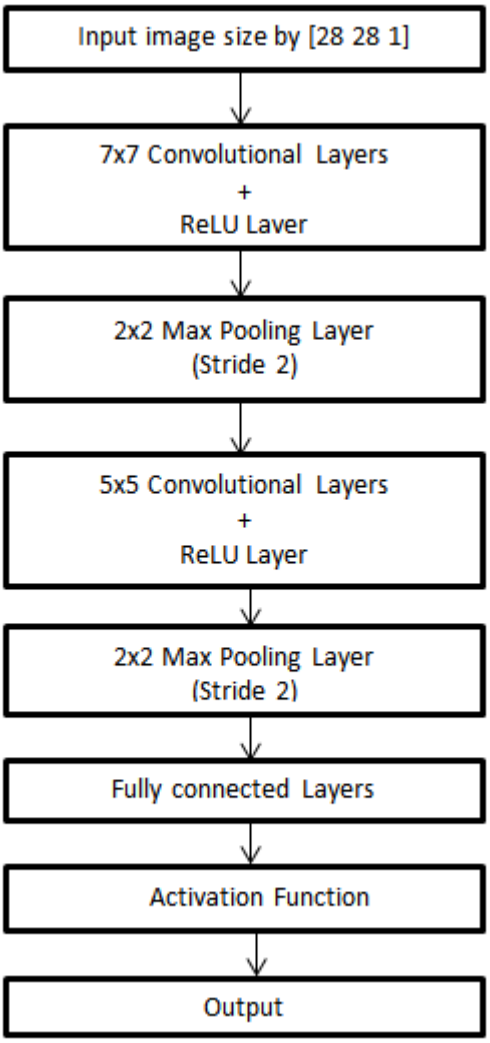


Figure 5.11: Block diagram of the CNN classification of harmful noise using two layers

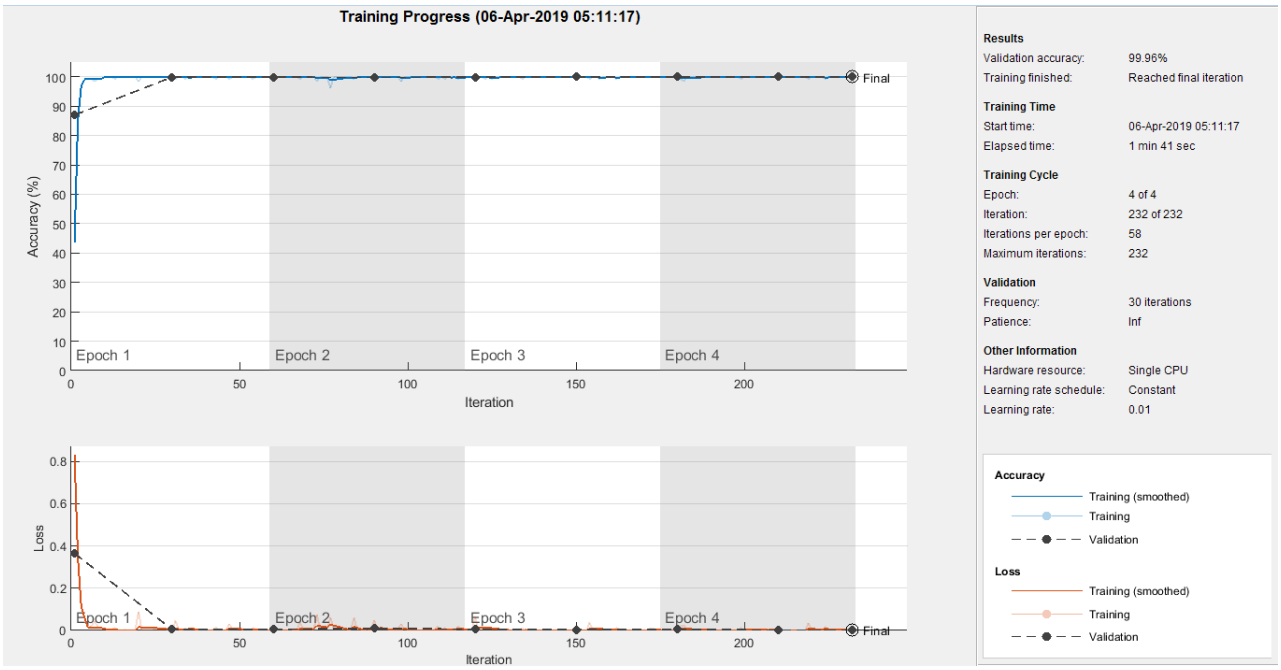


Figure 5.12: Validation accuracy of the speech and high frequency noise using two layers

5.5.3 CNN using Three Convolutional Layers

The block diagram of the CNN networks architecture is shown in figure 5.12 using three convolution2dLayer layers, two maxPooling2dLayers and one fully connected layer. The Figure 5.13 is the training progress plot. Here we use the first convolution2dLayer with filter size is 7x7 and channel size is 8 with the same padding, 2x2 maxPooling2dLayer with (stride2) .Then we apply second convolution2dLayer with filter size is 5x5 and channel size is 16 with the same padding, 2x2 maxPooling2dLayer with (stride2). Third convolution2dLayer with filter size is 3x3 and channel size is 32 with the same padding, 2x2 maxPooling2dLayer with (stride2) and fullyConnectedLayer is applied. After training the network the classification accuracy is 99.88%.

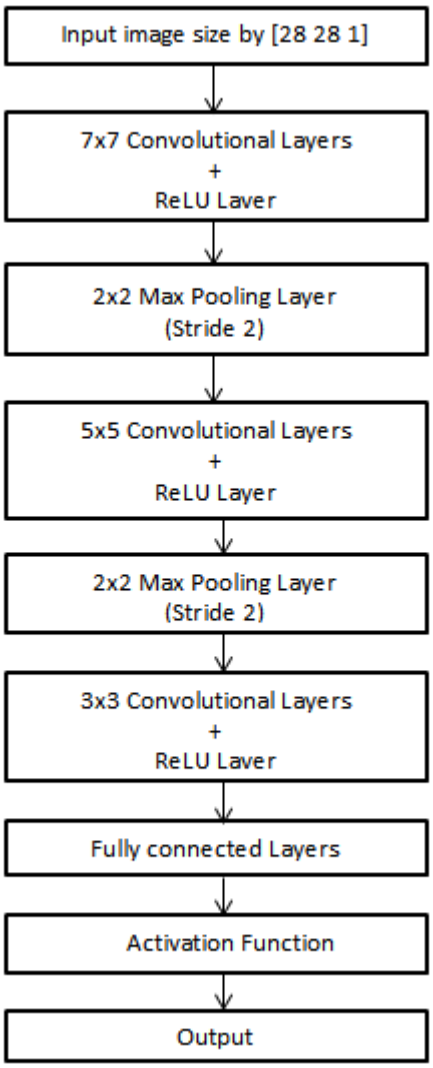


Figure 4.13: Block diagram of the CNN classification of harmful noise three players

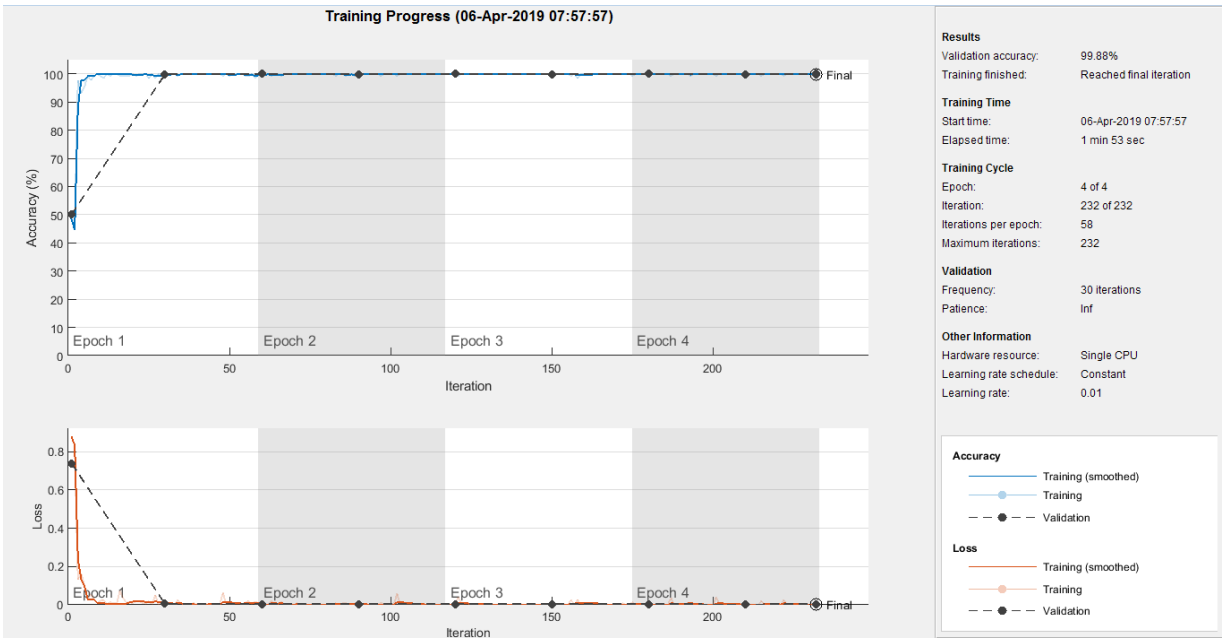


Figure 4.14: Evaluation accuracy of the speech and high frequency fixed noise using three layers

CHAPTER 6

EVALUATIONS

In evaluations, we use 10,000 spectrogram images in two classes. One class for clean speech which contains 5,000 spectrogram images and other class for noise images which contains 5000 spectrogram images. First we collected 100,000 recorded speech signals from (Mozilla Common Voice dataset. Speech signals durations are varying from 1 to 6 or 10 seconds. Then we convert these speeches to spectrogram images in Matlab by the spectrogram(x) function. During the training of CNN model, we use 3750 images of clean speech and 3750 images of speech+noise images for training. For validation 1250 images of clean speech and 1250 images of speech+noise images are utilized. After training a CNN model for each noise type (such as white noise, storm noise, bible noise, fixed noise and running tape noise) using 3750 clean speech and 3750 speech+noise of particular noise type. In particular, we obtain one CNN network for each noise type and use this network in the testing of a particular noise type.

6.1 Results - Data from the Same Dataset

The following tables and graphs show that when we train the network for a specific speech and noise (3750/3750 clean and speech+noise), then we test the network from seen images from the same dataset (1250/1250 clean and speech+noise). Results are shown in following tables and graphs

Table 6.1: Validation and test values of speech and noise signal (one convolutional layer)

Noise	Validation%	Test%
White Noise	100	100
Running Tape Noise	99.96	99.97
High Frequency Noise	99.88	99.85
Jet Aircrafts Noise	99.68	99.48
Storm Noise	99.48	99.64
Average	99.80	99.78

In table 6.1 for different noises, we only use one convolution layer with filter size [3 3] and channel size (neurons) is 8. For pooling size rectangular region is [2,2] is used with one maxpooling layer. The convolutional and down-sampling layers are followed by one or more fully connected layers. The output size is 2, corresponding to the 2 classes. Figure 6.1 shows the graphical representation of the validation and testing of speech and noise signals. validation means (training accuracy with the seen image samples in training) and test means means (testing with unseen images of 1250 from the same dataset).

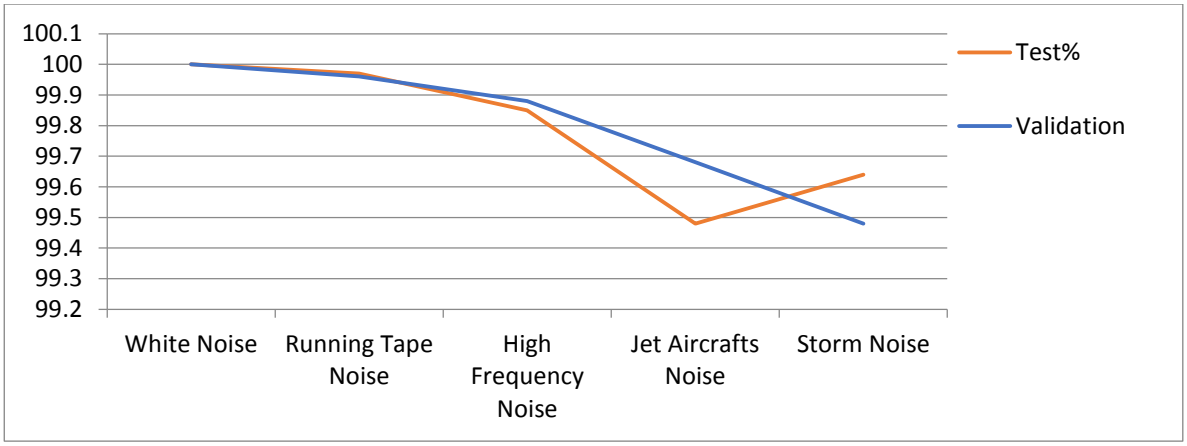


Figure 6.1: Validation and test values of speech and noise signal (one convolutional layer)

In table 6.2 for different noises we only use one convolution layer with filter size [3 3] and channel size (neurons) is 8. For pooling size rectangular region is [2,2] used with two maxpooling layers. The output fully connected layers size is 2. Figure 6.2 shows the results

Table 6.2: Validation and test values of speech and noise signal (two convolutional layer)

Noise	% Validation	Test%
White Noise	100	99.99
Running Tape Noise	99.96	99.99
High Frequency Noise	99.96	99.95
Jet Aircrafts Noise	99.80	99.61
Storm Noise	99.60	99.60
Average	99.80	99.79

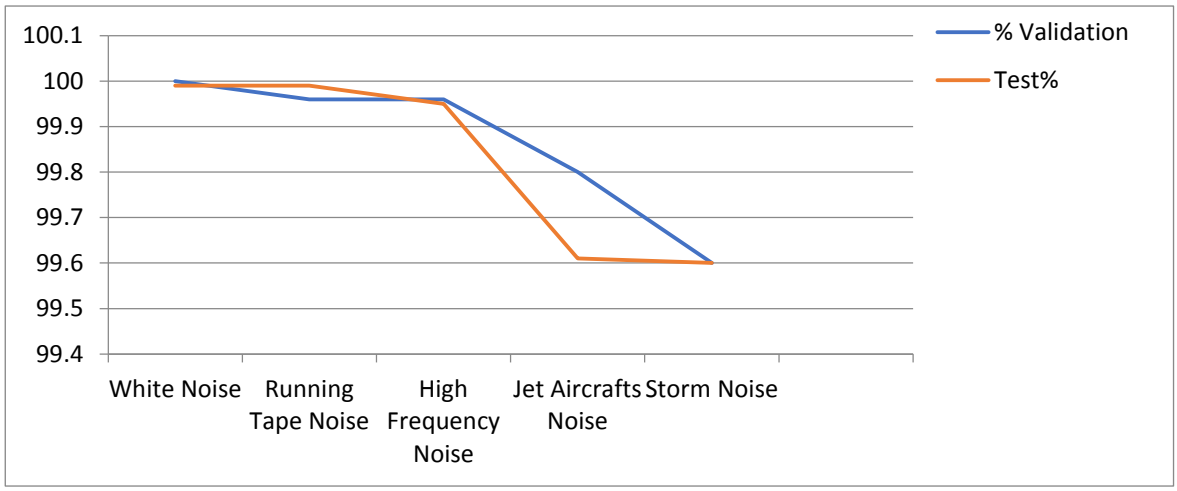


Figure 6.2: Validation and test values of speech and noise signals (one convolution layers)

In table 6.3 for different noises we use three convolution layers; convolutional layer one with filter size [3 3] and channel size (neurons) is 8. Second convolutional layer with filter size [3

3] and channel size (neurons) is 16 and third layer with filter size [3 3] and channel size (neurons) is 32. For pooling size rectangular region is [2,2] using two maxpooling layers. The output fully connected layers size is 2. Figure 6.3 shows the results.

Table 6.3: Validation and test values of speech and noise signal (three convolutional layers)

Noise	Validation %	Test%
White Noise	100	100
Running Tape Noise	100	100
High Frequency Noise	99.88	99.89
Jet Aircrafts Noise	99.52	99.75
Storm Noise	99.48	99.60
Average	99.77	99.84

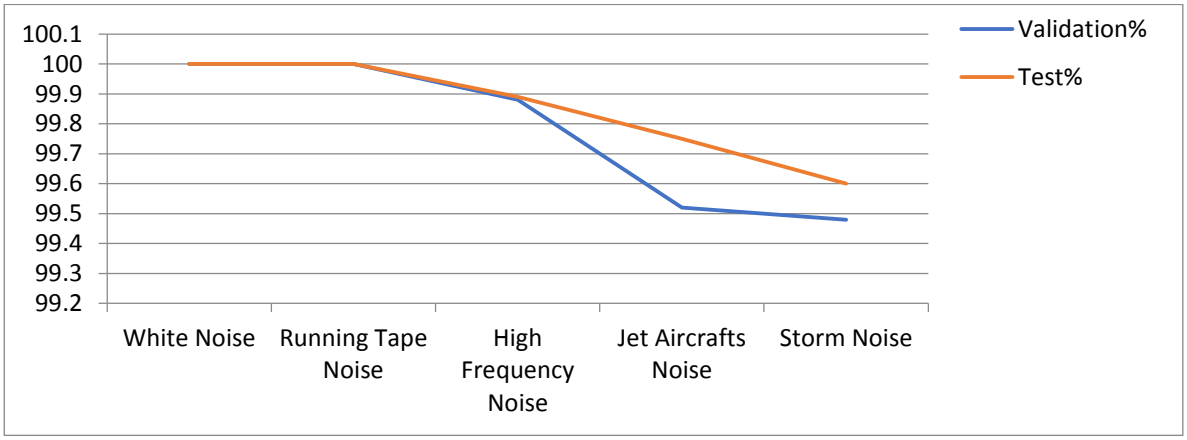


Figure 6.3: Validation and test values of speech and noise signal (three convolutional layers)

In table 6.4 for different noises, we only use one convolution layer with filter size [7 7] and channel size (neurons) is 8.For pooling size rectangular region is [2,2] used with one maxpooling layer. The output fully connected layers size is 2. Figure 6.4 shows the results.

Table 6.4: Validation and test of speech and noise signal (one convolutional layer)

Noise	Validation %	Test%
White Noise	100	99.97
Running Tape Noise	99.96	99.99
High Frequency Noise	99.92	99.89
Jet Aircrafts Noise	99.40	99.68
Storm Noise	99.32	99.55
Average	99.72	99.81

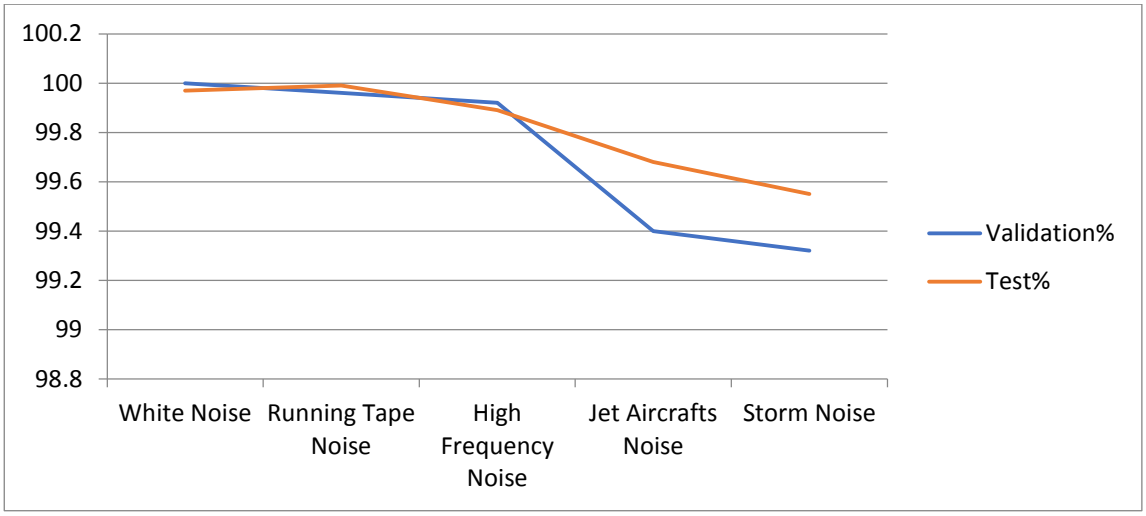


Figure 6.4: Validation and test values of speech and noise signal (one convolutional layer)

In table 6.5 for different noises, we use two convolution layers one with filter size [7 7] and channel size (neurons) is 8 and other layer with filter size [5 5] and channel size (neurons) is 16. For pooling size rectangular region is [2,2] used with two maxpooling layers. The output fully connected layers size is 2. Figure 6.5 shows the results.

Table 6.5: Validation and test of speech and noise signal (two convolutional layers)

Noise	Validation %	Test%
White Noise	99.96	99.92
Running Tape Noise	99.88	99.97
High Frequency Noise	99.88	99.93
Jet Aircrafts Noise	99.32	99.60
Storm Noise	99.60	99.44
Average	99.72	99.77

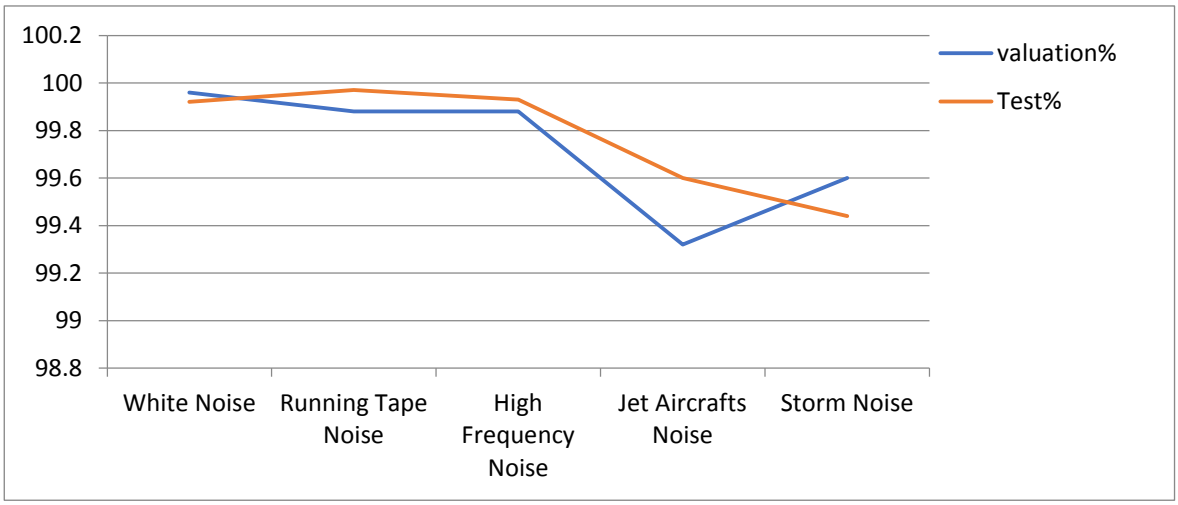


Figure 6.5: Validation and test values of speech and noise signals (two convolutional layers)

In table 6.6 for different noises we use three convolution layers; one with filter size [7 7] and channel size (neurons) is 8. Second layer with filter size [5 5] and channel size (neurons) is 16 and third layer with filter size [3 3] and channel size (neurons) is 32. For pooling size rectangular region is [2,2] used with two maxpooling layers. The output fully connected layers size is 2. Figure 6.6 shows the results.

Table 6.6: Validation and test values of speech and noise signal (three convolutional layers)

Noise	Validation %	Test%
White Noise	100	100
Running Tape Noise	99.96	99.99
High Frequency Noise	99.80	99.92
Jet Aircrafts Noise	99.36	99.65
Storm Noise	99.76	99.73
Average	99.77	99.85

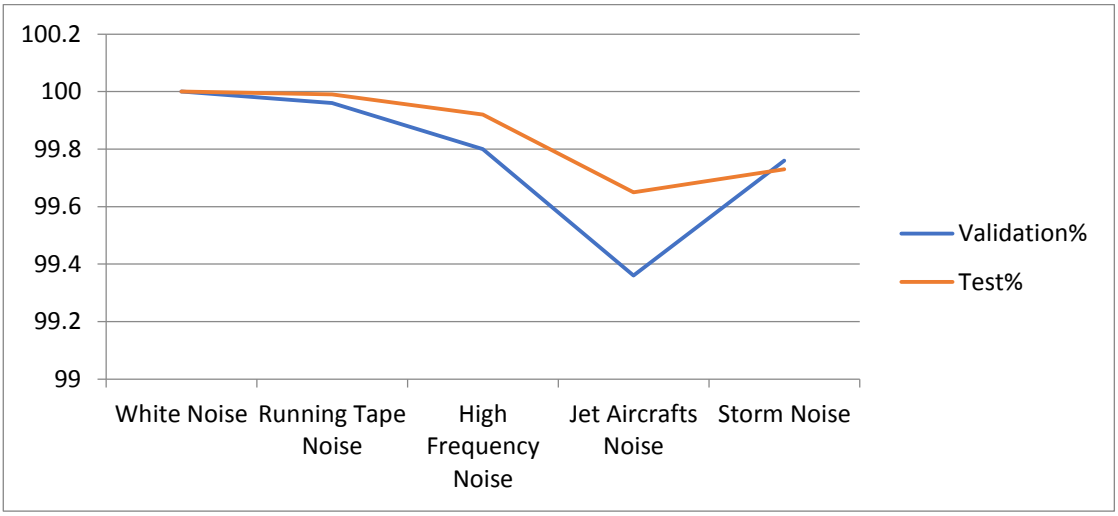


Figure 6.6: Validation and test values of speech and noise signals (threeconvolutional layers)

In all tables we compared speech and noise validation and test results for two-two classes.SoIt is clear CNN network results accuracy depend on the noise type.

6.2 Results – Data from Unseen Dataset

The following tables and graphs show that when we train the network for a specific speech and noise dataset and then test data from an unseen dataset, where CNN did not see (train) images from the same dataset. For this purpose, we downloaded speech signals from(Mozilla Common Voice dataset). In particular, we have 2000 testing speeches for each types of noise.

We converted these speeches to 2000 clean spectrogram images and added different noise types and obtain speech+noise spectrogram images. Results are shown in following tables and graphs.

The following tables and graphs show that when we train the network for a specific speech and noise (3750/3750 clean and speech+noise), then we test the network from unseen images from the same dataset (1250/1250 clean and speech+noise). Results are shown in following tables and graphs.

In table 6.7, for different noises we only use one convolution layer with filter size [3 3] and channel size (neurons) is 8.For pooling size rectangular region is [2,2] used with one maxpooling layer. The output fully connected layers size is 2, which corresponding to the 2 classes of the input data. Figure 6.7 shows the graphical representation of the validation and testing of speech and noise signals.

Table 6.7: Validation and test of speech and noise signal (one convolutional layer)

Noise	Validation %	Test%
White Noise	100	100
Running Tape Noise	99.96	100
High Frequency Noise	99.88	100
Jet Aircrafts Noise	99.68	99.72
Storm Noise	99.48	99.40
Average	99.80	99.83

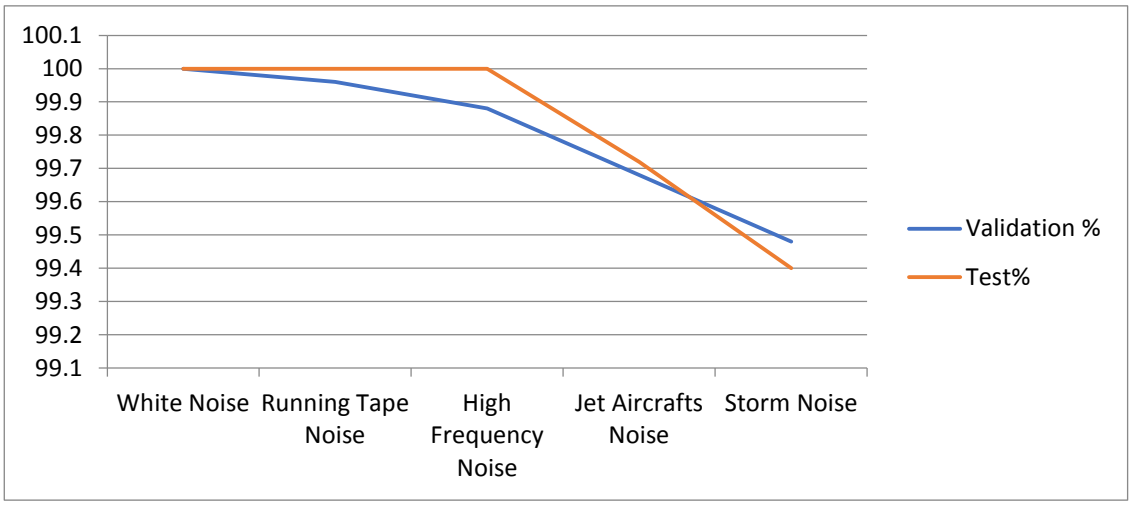


Figure 6.7: Validation and test of speech and noise signals (one convolutional layer)

In table 6.8 for different noises only use one convolution layers with filter size [3 3] and channel size (neurons) is 8.For pooling size rectangular region is [2,2] use two maxpooling layer. The output fully connected layers size is 2. Figure 6.8 shows the results.

Table 6.8: Validation and test of speech and noise signal (one convolutional layers)

Noise	Validation %	Test%
White Noise	100	100
Running Tape Noise	99.96	100
High Frequency Noise	99.96	100
Jet Aircrafts Noise	99.80	99.65
Storm Noise	99.60	99.45
Average	99.80	99.82

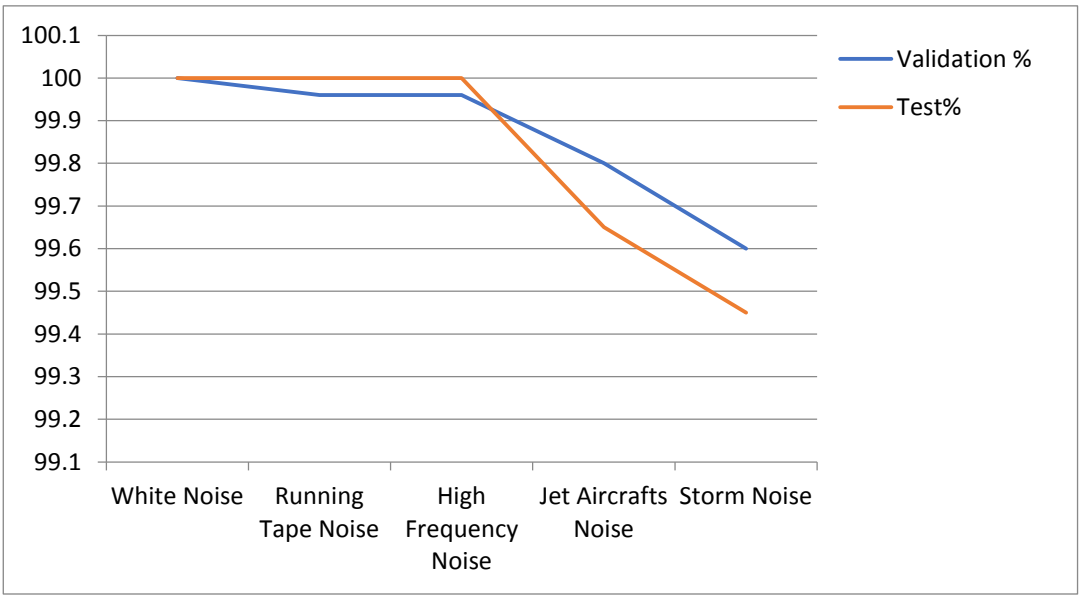


Figure 6.8: Validation and test of speech and noise signal (two convolutional layers)

In table 6.9 for different noises, we use two convolution layers; one with filter size [3 3] and channel size (neurons) is 8, second layer with filter size [3 3] and channel size (neurons) is 16 and third layer with filter size [3 3] and channel size (neurons) is 32.For pooling size rectangular region is [2,2] used with two maxpooling layers. The output fully connected layers size is 2. Figure 6.9 shows the results.

Table 6.9: Validation and test values of speech and noise signal (three convolutional layers)

Noise	Validation %	Test%
White Noise	100	100
Running Tape Noise	100	99.92
High Frequency Noise	99.88	99.85
Jet Aircrafts Noise	99.52	99.50
Storm Noise	99.48	99.60
Average	99.77	99.77

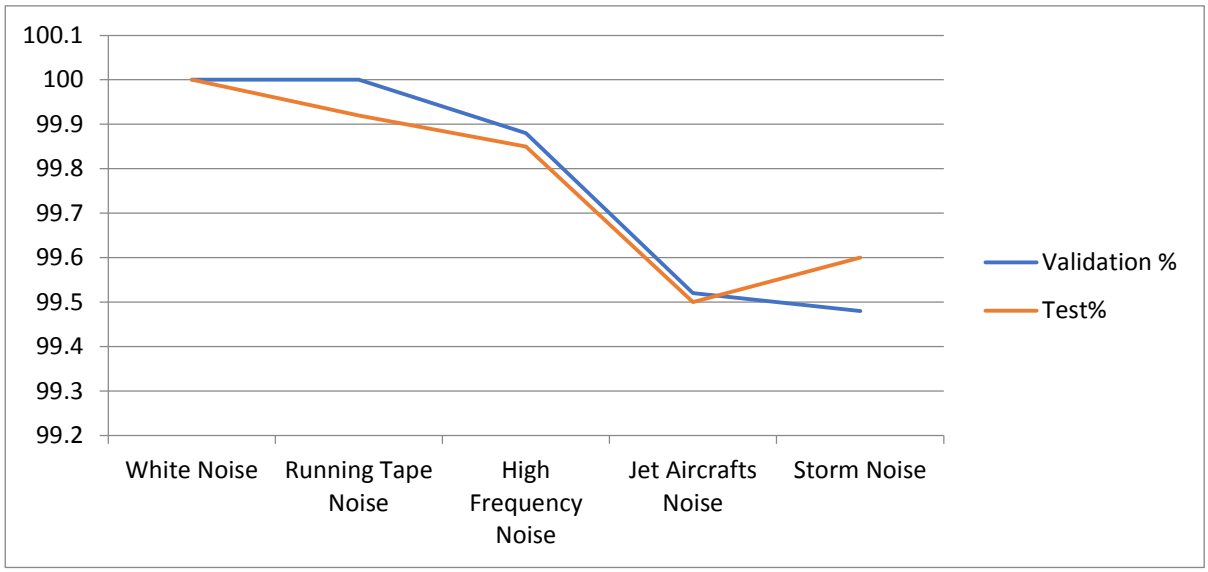


Figure 6.9: Validation and test values of speech and noise signal (three convolutional layers)

6.3 Results – Testing CNN network with different noise types

The following tables and graphs show that when we train the network for one kind of speech and noise signal and then test the network with different clean speech and speech+noise data on the network. We train 5000/5000 clean speech and speech+noise spectrogram images for a particular noise type from ... dataset, and then test 2000/2000 clean speech and speech+noise spectrogram images from the same dataset for a different noise type. Results are shown in following tables and graphs.

First we train a CNN network for white noise and check the validation for white noise and testing accuracy for other types of noises. In table 6.10 for different noises we only use one convolution layer with filter size [7 7] and channel size (neurons) is 8. For pooling size rectangular region is [2,2] used with one maxpooling layer. Figure 6.10 shows the graphical representation of the validation and testing of speech and noise signals. It shows in table White noise and running tape noise spectrograms are similar and CNN is able to classify running tape noise even it is trained with white noise. For other noise types it is random (around 50%).

Table 6.10: Validation and test values of speech and noise signals

Noise	Validation %	Test%
White Noise	100	100
Running Tape Noise		99.96
High Frequency Noise		50
Jet Aircrafts Noise		50
Storm Noise		49.96
Average		69.98

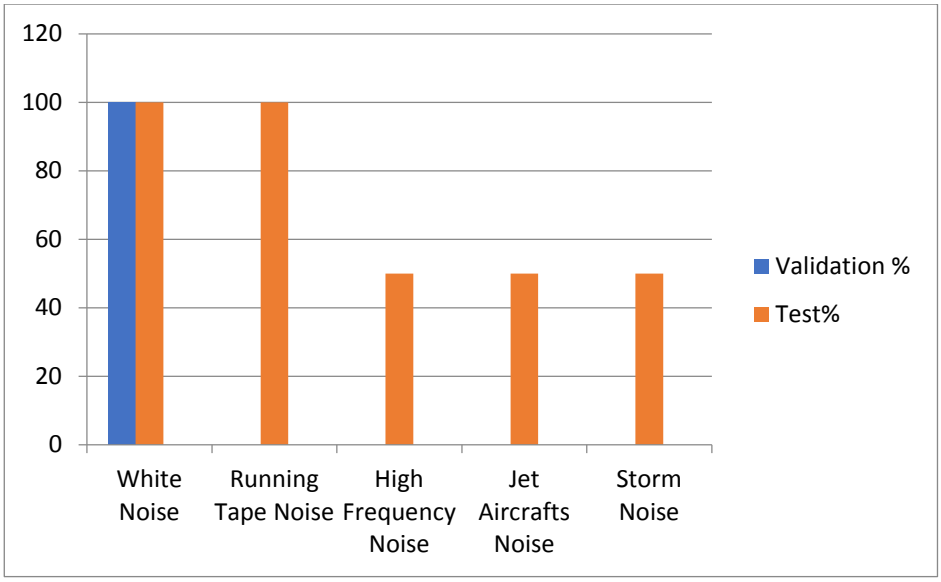


Figure 6.10: Validation and test of speech and noise signal.

We train a CNN network for running tape noise and we check the validation for running tape noise and testing accuracy for other types of noises. In table 6.11, for different noises we only use one convolution layer with filter size [7 7] and channel size (neurons) is 8. For pooling size rectangular region is [2,2] used with one maxpooling layer. Figure 6.11 shows the results. CNN is able to classify white noise even it is trained with running tape noise. For other noise types it is random (around 50%).

Table 6.11: Validation and test values of speech and noise signal

Noise	Validation %	Test%
White Noise		100
Running Tape Noise	99.6	99.96
High Frequency Noise		49.98
Jet Aircrafts Noise		50
Storm Noise		50
Average		69.98

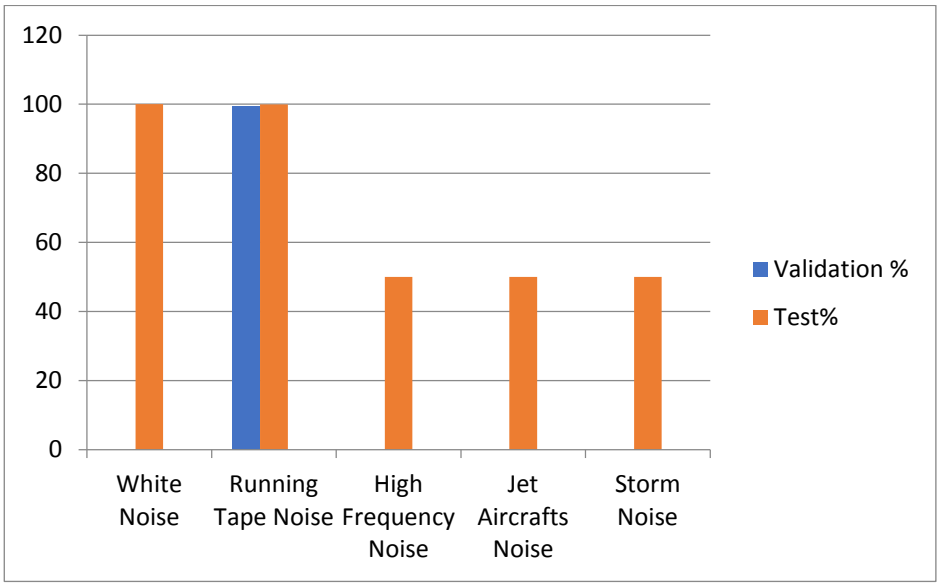


Figure 6.11: Validation and test values of speech and noise signals

We train a CNN network for high frequency noise and we check the validation for high frequency noise and testing accuracy for other types of noises. In table 6.12 for different noises, we only use one convolution layer with filter size [7 7] and channel size (neurons) is 8.For pooling size rectangular region is [2,2] used with one maxpooling layer. Figure 6.12 shows the results. It is shows in table White noise and high frequency noises spectrograms are similar. CNN is able to classify white noise even it is trained with high frequency noise. For other noise types it is random (around 50%).

Table 6.12: Validation and test values of speech and noise signal

Noise	Validation %	Test%
White Noise		99.90
Running Tape Noise		49.90
High Frequency Noise	99.92	100
Jet Aircrafts Noise		49.92
Storm Noise		50
Average		69.94

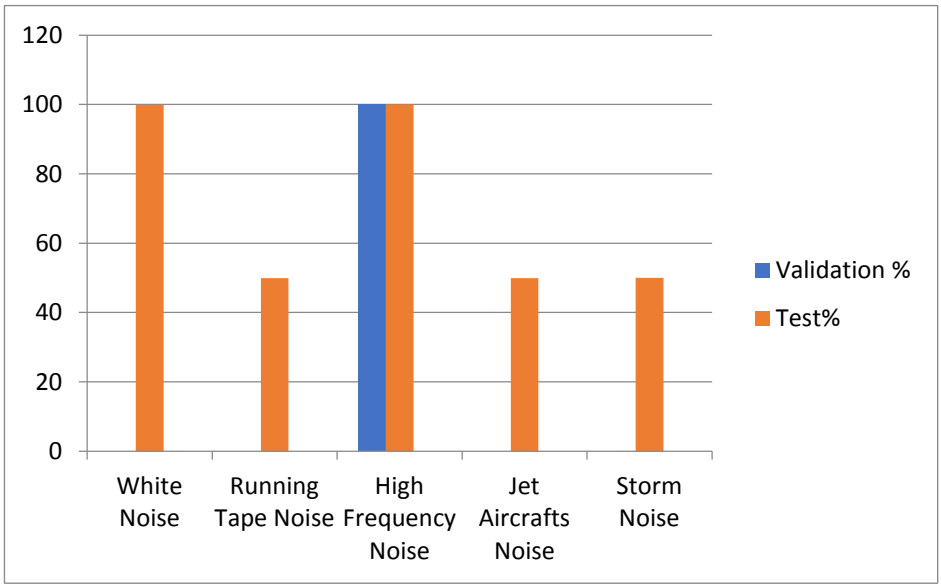


Figure 6.12: Validation and test values of speech and noise signals

We train CNN network for jet crafts noise and we check the validation for jet aircraft Noise noise and testing accuracy for other types of noises. In table 6.13, for different noises we only use one convolution layer with filter size [7 7] and channel size (neurons) is 8.For pooling size rectangular region is [2,2] used with one maxpooling layer. Figure 6.13 shows the results. White noise and jet craft noises spectrograms are similar. CNN is able to classify white noise even it is trained with high frequency noise. For other noise types it is random (around 50%).

Table 6.13: Validation and test values of speech and noise signals

Noise	Validation %	Test%
White Noise		99.76
Running Tape Noise		49.84
High Frequency Noise		50
Jet Aircrafts Noise	99.40	99.72
Storm Noise		49
Average		69.66

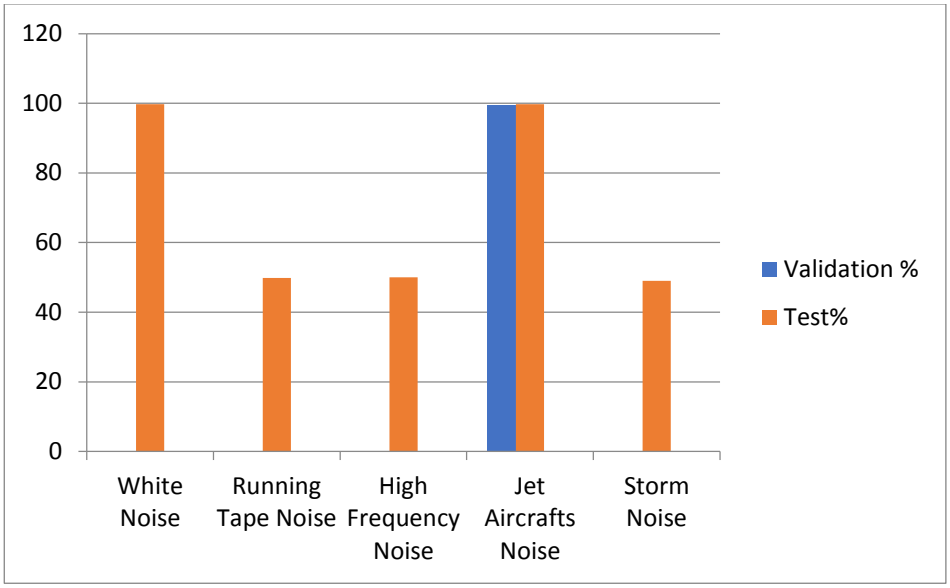


Figure 6.13: Validation and test values of speech and noise signal

We train a CNN network for storm noise and we check the validation for storm noise and testing accuracy for other types of noises. In table 6.14 for different noises we only use one convolution layer with filter size [7 7] and channel size (neurons) is 8. For pooling size rectangular region is [2,2] used with one maxpooling layer. Figure 6.14 shows the results.

Table 6.14: Validation and test values of speech and noise signal

Noise	Validation %	Test%
White Noise		50
Running Tape Noise		99.32
High Frequency Noise		49.80
Jet Aircrafts Noise		99.32
Storm Noise	99.32	99.40
Average		79.84

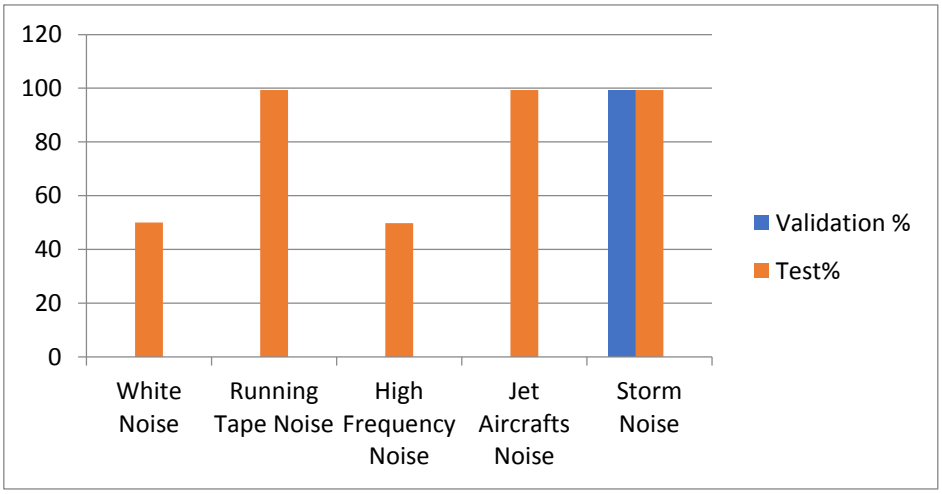
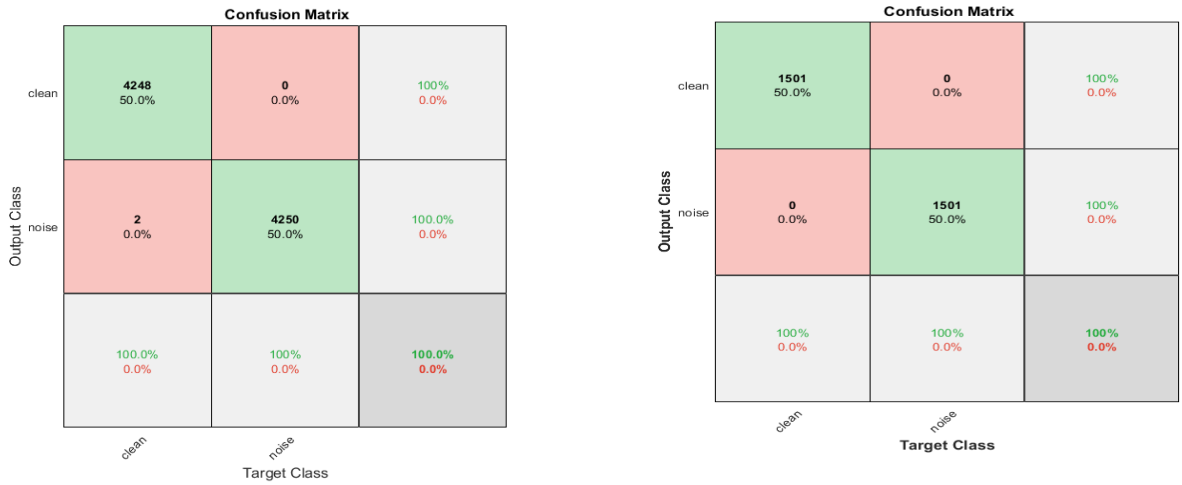


Figure 6.14: Validation and test values of speech and noise signal

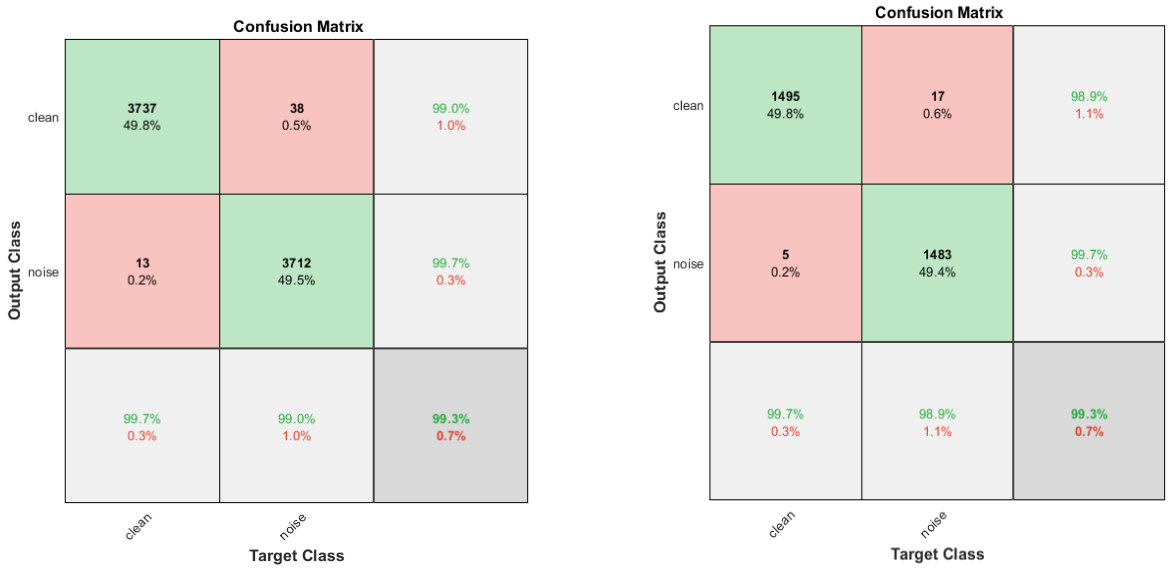
6.4 Confusion Matrix

Following is the confusion matrix of white noise and jet aircrafts. Figure 6.11 shows the in (a) the confusion matrix of the validation data while in (b) shows the confusion matrix of test data of the white noise. While Figure 6.12 shows the in (a) the confusion matrix of the validation data while in (b) shows the confusion matrix of test data of the jet aircrafts noise Figure 6.11(a) shows the confusion matrix trained to 5000 images. 4250 is for training and 750 for validation and Figure 6.11(b) which is the test confusion matrix of the train network.



(a) Confusion matrix for validation data (b) Confusion matrix for test data

Figure 6.15: Confusion matrix



(a) Confusion matrix for validation data (b) Confusion matrix for test data

Figure 6.16: Confusion matrix

6.5 Signals to Noise Ratio (SNR)

Signal to noise ratio (SNR) is the ratio of wanted signal (original signal) and unwanted signal(noise signal). When the SNR value is low its mean that the noise in the original signal is very high and when the SNR value is high its mean that the noise in the original signal is low. When the SNR value low, its mean that the spectrogram image has high noise then CNNs is able to classify harmful speech and un-harmful speech with high accuracy.so it is clear the classification accuracy of speech and noise signal depend on the SNR value.

SNR can be calculated by the following formula

SNR=Wanted Signal/Unwanted Signal

6.6 Epoch Table

The follow table show the different epochs results for high frequency noise. Epoch basically shuffle the data during train the network. While the training the network the number of epoch can affect the result of network which is explain in table.

Table No 6.15: Different epoch for fixed high frequency noise

Epoch	Validation	Test
1	99.92	99.92
2	99.88	99.83
4	99.92	99.92
6	99.84	99.92
8	99.96	99.96
10	99.92	99.96

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

Hearing loss is a major problem where many people in the world suffer from this disease. Therefore, hearing aid devices are very important for impaired people suffering from hearing loss. Analog hearing aid devices amplify the incoming sound wave. But they do not apply any techniques to reduce the noise of the speech signals. On the other hand, different people have various hearing problems such as some people have minor hearing loss, moderate hearing loss, severe hearing loss or profound hearing loss. As a result, special solutions are required such as digital hearing aid devices are adopted which are more flexible and able to adapt different hearing loss conditions compared to analog hearing aid devices. Generally these device use low-pass filter, adaptive filters and spectral analysis to remove harmful noise signal (i.e. airplane taking off, lightning), and amplify the incoming speech signal. However, in some cases, digital hearing aid devices cannot remove harmful noises, which may damage the ear. Therefore, this research focuses on classification of speech signals as harmful and unharmful using Convolution Neural Networks. In particular, first, we add different types of noises to the speech signal such as white noise, jet plane noise, storm noise, fixes frequency noise. Then, we try to remove these noises using different speech filters and analyze the speech. We observe that for some noise types, speech cannot be cleaned properly. To overcome this, we apply Convolution Neural Networks (CNN) to classify speech signals as harmful and unharmful in order to detect harmful speeches in digital hearing aid devices. To summarize, in this thesis, the following topics are studied:

- How analog hearing aids are working. Analog to the digital hearing aids are studied. Since digital hearing aids performance is better than analog hearing aids.
- Filters for noise reduction that are used in digital hearing aids are studied and implemented in Matlab.
- Convolutional Neural Networks (CNN) is studied. In particular, we converted speech signals to spectrogram images in order to classify with CNN.

- Several CNN architecture are designed to test the accuracy of different noise types for classification. In particular, the designed CNN networks can classify a harmful speech signal up to 99.99% accuracy. To the best of our knowledge no previous work has apply harmful unharmful speech classification using CNN using spectrogram images. Therefore, it is the contribution of the thesis.

In the current work, only speech signals are classified as harmful and unharmful. In future work, first we apply CNN for classification, then apply noise removal filter to the classified signal in order to remove the harmful speech signals. It is the future of digital hearing aids working together with CNN and digital filters such as adaptive filters.

REFERENCES

Amoh.J., & Odame, K. (2015, October). Deep Cough: A deep convolutional neural network in a wearable cough detection system. In 2015 IEEE biomedical Circuits and System Conference

Chang, S. Y., & Morgan, N. (2014). Robust CNN-based speech recognition with Gabor filter kernels. In *Fifteenth annual conference of the international speech communication association*.

Edwards,B. (2007). The futre of hearing aid technology. *Trends in amplification,11*(1).31-45.

Halawani, S. M., Al-Talhi, A. R., & Khan, A. W. (2013). Speech Enhancement Techniques for Hearing Impaired People: Digital Signal Processing based Approach. *Life Science Journal, 10*(4).

Hinton, G. E.,., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feed forward neural networks. In*Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256).

Jati, A., & Georgiou, P. (2018). Neural predictive coding using convolutional neural networks towards unsupervised learning of speaker characteristics. *arXiv preprint arXiv:1802.07860*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing system* (pp. 1097-1105).

Levitt, H. (2001). Noise reduction in hearing aids: A review.*Journal of rehabilitation researh and development development*, *38*(1), 111-122.

Lebart, K., Boucher, J. M., & Denbigh, P. N. (2001). A new method based on spectral subtraction for speech dereverberation. *Acta Acustica united with Acustica*, *87*(3), 359-366

Lin, C. H., Tsai, S. H., & Chuang, G. C. (2013, May). A novel Sub-Nyquist sampling of sparse wideband signals. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4628-4632). IEEE.

Li, X., & Zhou, Z. Speech Command Recognition with Convolutional Neural Network.

Magotra, N. (2000, August). Development of a low power digital hearing processor. In *Int'l Hearing Aid Research Conf*.

Ngo, K. (2011). *Digital signal processing algorithms for noise reduction, dynamic range compression, and feedback cancellation in hearing aids* (Doctoral dissertation, PhD thesis,ESATKatholieke Universiteit Leuven, Belgium).

Palaz, D., & Collobert, R. (2015). *Analysis of cnn-based speech recognition system using raw speech as input* (No. REP_WORK).

Piczak, K. J. (2015, September). Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1-6). IEEE.

Ramírez Frías, A. (2018). Applications of deep learning in the analysis of medical images.

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *arXiv gpreprint arXiv:1710.05941*.

Rodman, J.(2003). The effect of bandwidth on speech intelligibility. *White paper, polycom,Inc*

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211-252.

Sailor, H. B., Agrawal, D. M., & Patil, H. A. (2017, August). Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification. In *INTERSPEECH* (pp. 3107-3111).

Tanyer, S. G., & Ozer, H. (2000). Voice activity detection in nonstationary noise. *IEEE Transactions on speech and audio processing*, *8*(4), 478-482.

APPENDIX

```
% cd C:\Users\Khalid Zaman\Downloads\cv_corpus_v1\HF
cd 'D:\5000\whitenoise Train'

% matlabroot='C:\Users\Zaman\Downloads\cv_corpus_v1\';
matlabroot='D:\5000\';
digitDatasetPath = fullfile(matlabroot,'whitenoise Train');
imds = imageDatastore(digitDatasetPath, ...
   'IncludeSubfolders',true,'LabelSource','foldernames');

figure;
perm = randperm(20,20);
for i = 1:20
   subplot(4,5,i);
   imshow(imds.Files{perm(i)})
end
labelCount = countLabel(imds);

img = readimage(imds,1);
size(img)
numTrainFiles = 1250;
[imdsTrain,imdsValidation] = splitEach(imds,numTrainFiles,'randomize')

layers = [
   imageInputLayer([28 28 1])

   convolutiondLayer(3,8,'Padding','same')
   batchNormalizationLayer
   reluLayer

   maxPooling2dLayer(2,'Stride',2)

   convolutiondLayer(3,16,'Padding','same')
   batchNormalizationLayer
   reluLayer

   maxPoolingdLayer(2,'Stride',2)

   convolutiondLayer(3,32,'Padding','same')
   batchNormalizationLayer
```

```
    reluLayer
    fullyConnectedLayer(2)
    softmaxLayer
    classificationLayer];


% options = trainingOptions('sgdm', ...
%     'InitialLearnRate',0.01, ...
%     'MaxEpochs',8, ...
%     'Shuffle','every-epoch', ...
%     'ValidationData',imdsValidation, ...
%     'ValidationFrequency',30, ...
%     'Verbose',false);


options = trainingOptions('sgdm', ...
    'InitialLearnRate',0.01, ...
    'MaxEpochs',4, ...
    'Shuffle','every-epoch', ...
    'ValidationData',imdsValidation, ...
    'ValidationFrequency',30, ...
    'Verbose',false, ...
    'Plots','training-progress')

% options = trainingOptions('sgdm')

net = trainNetwork(imdsTrain,layers,options)
cd D:\5000
Gnet66=net;
save Gnet66;


% cd C:\Users\Zaman\Downloads\cv_corpus_v1\
% load Gnet2;
% net=Gnet2;

YPred = classify(net,imdsValidation);
YValidation = imdsValidation.Labels;
plotconfusion(YValidation,YPred)
accuracy = sum(YPred == YValidation)/numel(YValidation)

% cd C:\Users\Zaman\Downloads\cv_corpus_v1\T2HF
cd 'D:\5000\white2000'
% matlabroot='C:\Users\Zaman\Downloads\cv_corpus_v1\';
matlabroot='D:\5000\';
digitDatasetPath = fullfile(matlabroot,'white2000');
```

```
imds = imageDatastore(digitDatasetPath, ...
   'IncludeSubfolders',true,'LabelSource','foldernames');

figure;
perm = randperm(1000,20);
for i = 1:20
   subplot(4,5,i);
   imshow(imds.Files{perm(i)})
end
labelCount = countEachLabel(imds)

img = readimage(imds,1);
numTrainFiles =500;



[ximdsTrain,ximdsValidation] = splitEachLabel(imds,numTrainFiles,'randomize')
YPred = classify(net,ximdsValidation);
xValidation = ximdsValidation.Labels;
plotconfusion(xValidation,YPred)
[ximdsValidation(1:10,:) YPred(1:10,:)];
xValidation = ximdsValidation.Labels;
net = trainNetwork(ximdsTrain,layers,options)

Test_accuracy = sum(YPred == xValidation)/numel(xValidation)


Fs = 44100;

y = wavrecord (3*Fs,Fs);

wavplay (y,Fs);

c =fft(y);

subplot(3,1,1) ,plot(abs(c),'b');

xlabel('(Frequency)')

ylabel('Amplitude');

legend('Orignal Signal ');

y=awgn(y,30);

wavplay (y,Fs);

d= fft(y);

subplot(3,1,2) ,plot(abs(d),'g');

xlabel('(Frequency)')
```

```
ylabel('Amplitude');
legend('Signal+Noise ');
[b,a] = cheby1(5,1,[0.04 0.07]);
Hd = dfilt.df2t(b,a);
y = filter(Hd,y);
wavplay(y,Fs);
e= fft(y);
subplot(3,1,3),plot(abs(e),'r')
xlabel('(Frequency)')
ylabel('Amplitude');
legend('filter signal');


Fs=44100;
y1=audioread('sound888.mp3');
sound(y1,Fs)
subplot(4,1,1) ,plot(y1,'b');
xlabel('(Time)')
ylabel('Amplitude');
legend('Normal Speech');
y2=audioread('f16.mp3');
sound(y2,Fs)
subplot(4,1,2) ,plot(y2,'r');
xlabel('(Time)')
ylabel('Amplitude');
legend('High Frequency Noise');
minimumlength=min([length(y1),length(y2)]);
y1=y1(1:minimumlength);
y2=y2(1:minimumlength);

y3=y1+y2;
sound(y3,Fs)
subplot(4,1,3) ,plot(y3,'g');
xlabel('(Time)')
ylabel('Amplitude');
legend('Speech+High Frequency Noise');

b = fir1(31,0.5);
d = filter(b,1,y1)+y3;
mu = 0.001; %amplitude
ha = adaptfilt.lms(32,mu);
[y,e] = filter(ha,y1,d);
```

```
sound(y,Fs)
subplot(4,1,4) ,plot(y);
xlabel('(Time)')
ylabel('Amplitude');
legend('Speech by Adaptive Filter');


Fs = 44100;
y = wavrecord (3*Fs,Fs);
wavplay (y,Fs);
subplot(3,1,1) ,plot(y,'b')
xlabel('(Frequency)')
ylabel('Amplitude');
legend('Orignal Signal ');
y=awgn(y,30);
subplot(3,1,2) ,plot(y,'g');
xlabel('(Frequency)')
ylabel('Amplitude');
legend('Signal+Noise ');
[b,a] = butter(6,.02);
Hd = dfilt.df2t(b,a);
y = filter(Hd,y);
subplot(3,1,3),plot(y,'r')
xlabel('(Frequency)')
ylabel('Amplitude');
legend('filter signal')
```