TECHNISCHE UNIVERSITÄT MÜNCHEN

PROJECT REPORT

Source Separation based on Deep Learning

Author:
Tapan SHARMA

Supervisor:
Ms. Han Li

Examiner:
Prof. Dr.-Ing. Bernhard U. Seeber

Chair of Audio Signal Processing
Faculty of Electrical Engineering and Information Technology

July 17, 2019

Declaration of Originality

This is to certify that the project titled **"Source Separation based on Deep Learning"** is my original work and is being submitted to the chair of the "**Audio Information Processing**" of the "**Faculty of Electrical and Information Technology**" in the "**Technische Universität München**". This report has not been submitted earlier either to this University or to any other University/Institution for the fulfillment of the requirement of a course of study.

Name: **Tapan Sharma**
Place: **Munich**
Date: **12th July, 2019**

Acknowledgements

I would first of all like to thank the "Chair of Audio Information Processing" under the leadership of Prof. Dr.-Ing. Bernhard U. Seeber for making this project lab course available to the masters students as a module particularly in the curriculum program of MSCE (Master of Science in Communication Engineering).

I would also like to thank our project supervisor Ms. Han Li who structured and organised the problem statement for the project very clearly and also motivated us to take on the project work with well defined goals to achieve. Her guidance has been in particular of great help to me in improving my work. She pointed us to the right direction in terms of articles, websites and toolboxes which turned out to be of great help to us.

In the end, this project would not have seen timely completion without the help of my project mate Mr. Md. Toaha Umar. His insights in our discussions were very fruitful and helped me deepen my understanding and get a new perspective to approach solving some problems. In all, working with him has helped me a lot in my peer learning.

Abstract

This report presents an implementation and evaluation of an end-to-end system for speech enhancement based source separation, in the monaural voice recordings. Speech Enhancement is achieved by applying a ratio mask to a time-frequency representation of the input signal and through a subsequent reconstruction for the estimated clean speech signal. The mask is estimated from the noisy mixture input data using deep learning machines which are trained on a dataset obtained by additive mixing of recordings of the clean speech and the different sources of noise at a particular signal-to-noise ratio. The expected intelligibility of the reconstructed audio is compared using the intelligibility metric STOI (Short-Time Objective Intelligebility). The quality of estimated clean speech post processing, is compared using the quality metric PESQ (Perceptual Evaluation of Speech Quality), as recommended by the ITU (International Telecommunication Union). We discover that a deep convolutional neural network based learning machines, achieve a superior performance at speech enhancement using the cochleagram as the training data along with the time-frequency masks based on cochleagrams when compared to other deep learning models working on acoustic features and time-frequency masks as their training data.

Contents

List of Figures

List of Tables

List of Abbreviations

STFT	**S**hort-**T**ime **F**ourier **T**ransform
ERB	**E**quivalent **R**ectangular **B**andwidth
STOI	**S**hort-**T**ime **O**bjective **I**ntelligibility
PESQ	**P**erceptual **E**valuation of **S**peech **Q**uality
MOS	**M**ean **O**pinion **S**core
LQO	**L**istening **Q**uality **O**bjective
PESQMOS	**P**erceptual **E**valuation of **S**peech **Q**uality **M**ean **O**pinion **S**core
MOSLQO	**M**ean **O**pinion **S**core **L**istening **Q**uality **O**bjective
DNN	**D**eep **N**eural **N**etwork
CNN	**C**onvolutional **N**eural **N**etwork
LSTM	**L**ong **S**hort-**T**erm **M**emory
IRM	**I**deal **R**atio **M**ask
GFCC	**G**ammatone **F**requency **C**epstral **C**oefficients
MFCC	**M**el-**F**requency **C**epstral **C**oefficients
t-f	**t**ime-**f**requency
RMSE	**R**oot **M**ean **S**quare **E**rror
reLU	**re**ctified **L**inear **U**nit

List of Symbols

t	time	s
f	frequency	Hz
$S(t,f)$	log spectral energy of clean speech signal	dB per t-f frame
$N(t,f)$	log spectral energy of noise signal	dB per t-f frame

Chapter 1

Introduction

1.1 Why Source Separation?

For a healthy person it is an incredible feat to tune into a single conversation in a crowded room while being completely oblivious to background noise and other conversations that happen simultaneously. This ability for a person to solve the "Cocktail Party Problem"[1] is remarkable and very natural as a human. However, an estimated 5% of the world's population suffers from disabling hearing loss [2]. For the 466 million affected people the ability to perceive speech especially in the presence of background noise is significantly impaired. Since the perception of speech is crucial for being able to communicate verbally, this impairment has profound social and emotional consequences for the sufferers [2].

Current hearing aids offer only limited relief as they cannot distinguish between speech and background noise and will simply opt to amplify both signals. This does not greatly help with the intelligibility. Hence, the focus should be on designing an "intelligent" system that can distinguish between the conversation and disruptive background noise in order to provide an improvement to hearing aids. It is not unconceivable that such systems would also be of great help to the people with a normal and healthy hearing considering the fact that listening to someone in a loud environment is still detrimental to the hearing.

An estimated half of all cases of hearing loss in adults are caused by exposure to excessively loud noise [3]. A system that can remove loud background noise but allow for unhindered conversation might therefore help prevent such cases of hearing loss.

This project report contributes to these goals by comparing different implementations of supervised deep learning systems that can learn to enhance speech from background noise in monaural (single microphone) recordings.

1.2 Historical Work

The first approach to solve the Cocktail Party Problem was using the concept of "**Spectral Subtraction**"[4]. This approach makes an assumption that the clean speech signal has been corrupted by statistically independent additive noise. The power spectrum of the noise signal can be recovered approximately for stationary noise by taking the average of multiple signal frames. This estimated power spectrum is then subtracted from the mixed signal spectrum, keeping the phase information intact. The resulting subtraction results in less noise in the estimated signal but the quality of such estimation also tends to deteriorate considering the fact that many tones tend to appear and disappear rapidly in the reconstructed signal.

The idea to use a deep learning machine in order to filter out background noise from speech recordings goes back three decades. In 1988 Tamura and Waibel [5] showed that a four-layer feed-forward neural network could successfully be used to directly separate the waveforms of Japanese speech recordings from computer lab background noise. This network is small by the standards of today, but took weeks of training on a supercomputer of its time. The authors note that the perceived quality of the filtered audio was higher, but that intelligiblity did not improve.

Artificially adding noise to existing audio recordings is routinely used in supervised training of automatic speech recognition systems in order to improve their robustness towards environmental noise [6]. Chen, Jitong and Wang [7] follow an approach rooted in the computational auditory scene analysis (CASA). The idea is to predict an "Ideal Ratio Mask" (IRM) that describes the proportion of speech to noise energy in a frequency band at a given time. When the IRM is known, the clean speech signal can be reconstructed from the noisy signal by weighting each time frequency unit of the noisy recording according to its speech content.

Huang, Kim, Hasegawa-Johnson and Smargdis further improve this approach by presenting a framework for separating arbitrarily many sources using recurrent neural networks (RNNs) using a discriminative training criterion[8]. They also move the time-frequency masking operation directly into a layer of the neural network in order to jointly optimize the network with the masking function.

1.3 Limitations

There are several assumptions made in this project which limit the direct application of the results in a real time situations. These limitations can however be overcome by acquiring more training data.

- While the recordings of noise included complex interactions of the sound sources with the environment like factory noise, waterfalls, engine noise, etc., the speech examples were considered from relataively simpler scenarios without any interaction of sound sources with the real world conditions. Hence, many effects like reverberated speech are not modeled.

- *Lombard Effect [9]*: It is the ability of humans to alter their speaking style in order to allow for efficient verbal communication in a noisy environment. This too has not been considered. Hence, there is likely to be a mismatch between the prediction of the data considered in the project with the real world data.

Chapter 2

Methodology

2.1 Supervised learning architecture for speech enhancement

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs [10]. In supervised learning, each example is a pair consisting of an input object and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.

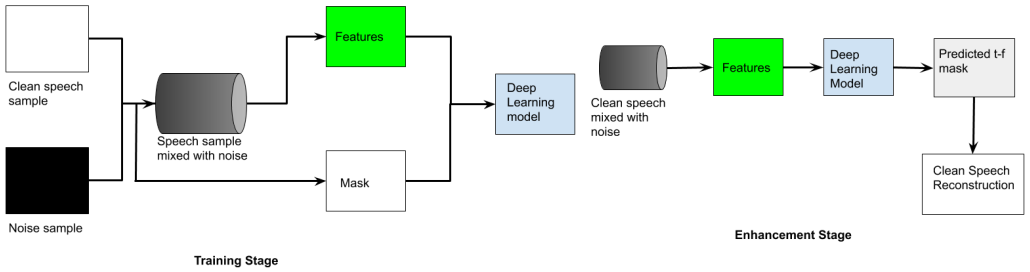
The idea of using supervised learning for speech-enhancement involves using the training data consisting of audio features as input object and the desired t-f mask as an output value during the training phase (See figure: 2.1a). This would enable a deep learning model to learn a function to model the training data. The trained deep learning model can then be used on a testing dataset that would only involve audio features as input to the trained deep learning model. Post training phase, the deep learning model would then predict the t-f mask for the input audio features (See figure: 2.1b).

2.2 Dataset

Dataset forms the most important part of any supervised learning problem. For the experiments in this project, only the speech portion, which consists of read speech (TIMIT recordings), and the set of noises(MUSAN recordings), which ranges from beeps emitted from technical equipment, to ambient sounds such as rain, road , factory noise, etc) were considered. All of the files are available in ".wav" format, are single channel, and are 16 bit sample PCM encoded. All recordings are downsampled to 8kHz sampling rate to speed up the training phase.

- **Clean Speech Dataset**:
  TIMIT: This corpus [11] consists of a total of 1700 sentences spoken by roughly 630 speakers of eight major dialects of American English. TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance. To speed up the training in our case however, these samples were downsampled to 8kHz. Corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments, Inc. (TI). The speech was recorded at TI, transcribed at MIT and verified by the National Institute of Standards and Technology (NIST).

- **Noise Dataset**:
  MUSAN: This is a corpus of music, speech, and noise recordings [12]. This



(A) Training Stage Layout (B) Testing Stage Layout

FIGURE 2.1: Deep Learning based speech enhancement layouts

dataset consists of music, speech recordings in twelve different languages, and a large set of naturally occuring and technical noises. Its primary intention is that of being a training corpus for voice activity detection. We considered the noise sub set of this corpus which amounts to around 6 hours of the duration.

2.2.1 Generation of Noisy Mixture

During training phase, the clean speech sample was mixed with a randomly chosen noise sample after making both samples of same duration and at the same amplitude level, i.e. both clean speech and noise samples were first normalised to have same amplitude and then added together. The mixing procedure has been depicted as a flow chart in the figure 2.2.

2.3 Features

In machine learning and pattern recognition, a feature is an individual measurable property or characteristic of a phenomenon being observed [13]. Choosing informative, discriminating and independent features is a crucial step for effective algorithms in pattern recognition, classification and regression.

Features as input and learning machines play complementary roles in supervised learning. When features are discriminative,they place less demand on the learning machine in order to perform a task successfully. On the other hand, a powerful learning machine places less demand on features.

We conducted a study to examine different acoustic features for supervised speech enhancement:

2.3.1 Spectrogram

A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. Spectrogram for an audio is constructed by combining the STFT at subband levels. Audio samples at 8 KHz are divided into frames of 30 ms. Hence, each frame has 240 samples for which STFT is calculated. An overlap of 60 samples

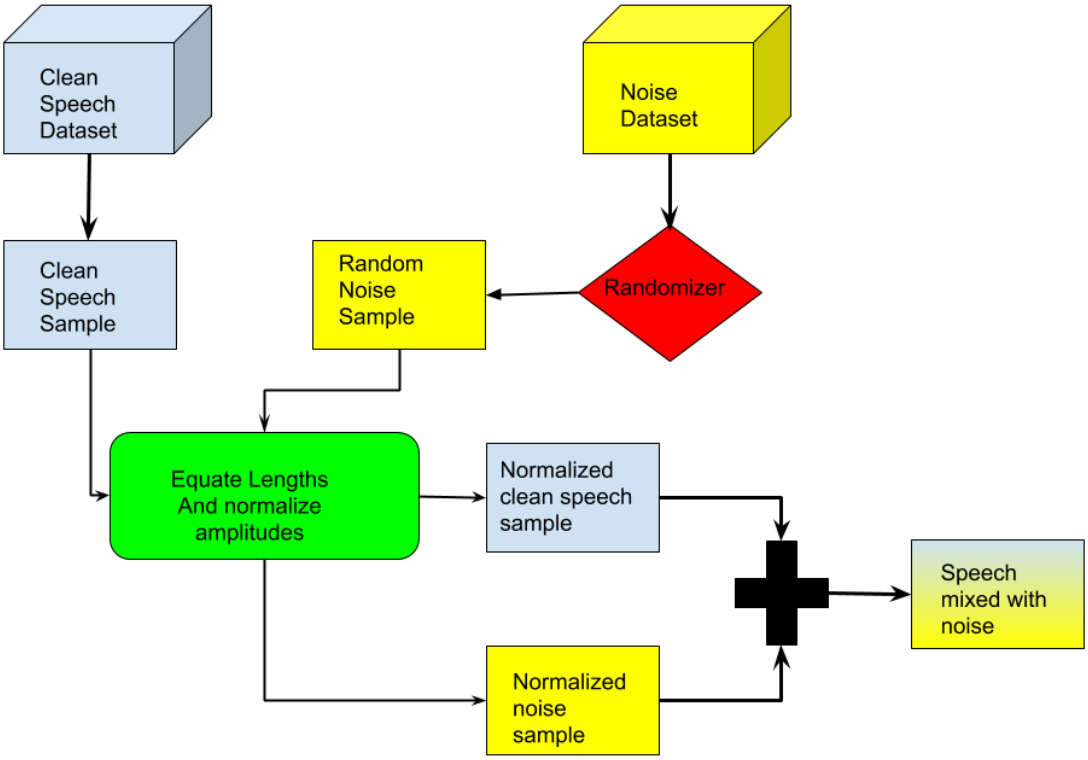
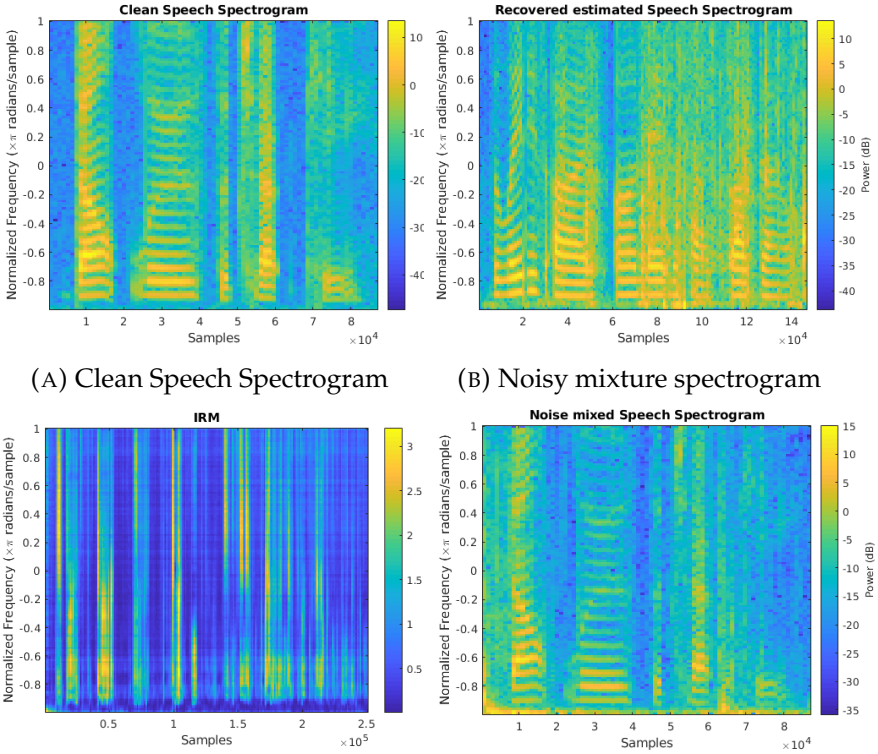


FIGURE 2.2: Generation of Noisy speech samples



(A) Clean Speech Spectrogram (B) Noisy mixture spectrogram

(C) IRM - by learning model (D) Estimated speech spectrogram

FIGURE 2.3: Different stages of spectrogram during speech enhancement

in adjacent frames is also used to reduce the effect of windowing. Combining STFT for all subband frames gives a spectrogram which can then be used as the training data and for generating spectrogram masks (See figure: 2.3).

2.3.2 GFCC

GFCC (Gammatone Frequency Cepstral Coefficients) are the gammatone-domain cepstrum based audio feaures. Being a cepstral property, it is a spectral representation of the spectrum of a time domain audio signal which has been filtered using a gammatone filterbank. These features are calculated at subband levels for frames of length 30ms. For an audio sampled at 8KHz rate, this imply 240 samples per frame. To reduce the effect of windowing, an overlap of 60 samples in adjacent frames is considered. This method to calculate GFCC, the rate of change of GFCC called GFCC delta and rate of change of GFCC delta, called GFCC double delta is described in 2.5. In all 13 gammatone frequency cepstral coefficients were considered (See figure: 2.4 [14]).

2.3.3 MFCC

MFCC (Mel-Frequency Cepstral Coefficients) are the Mel-domain cepstrum based audio features. It is a spectral representation of the spectrum of a time domain audio signal which has been filtered using a mel-frequency filterbank. The MFCC is calculated by splitting the entire data into overlapping segments. These features are calculated at subband levels for frames of length 30ms. For an audio sampled at 8KHz rate, this imply 240 samples per frame. To reduce the effect of windowing, an

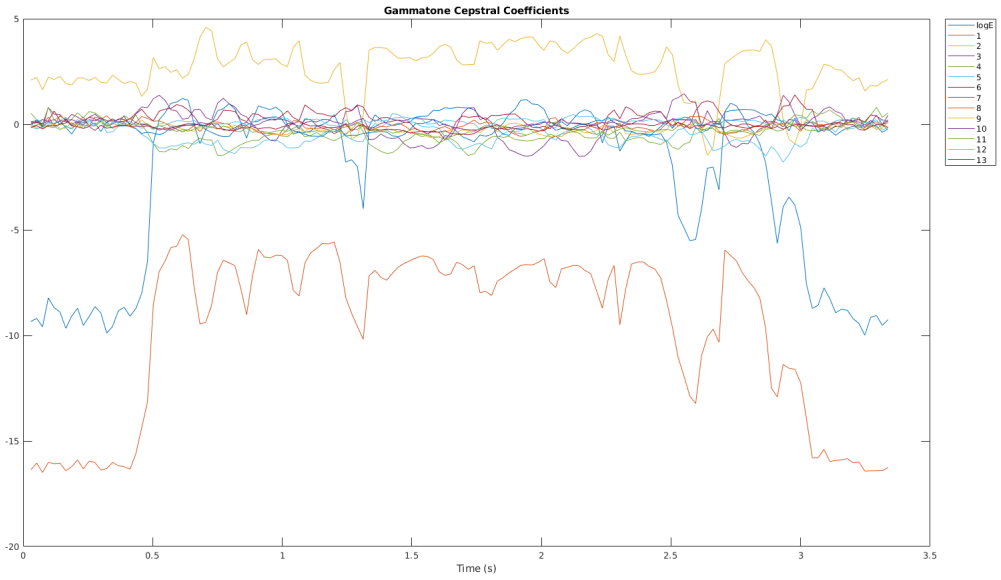


FIGURE 2.4: Typical GFCC plot with log energy and 13 coefficients

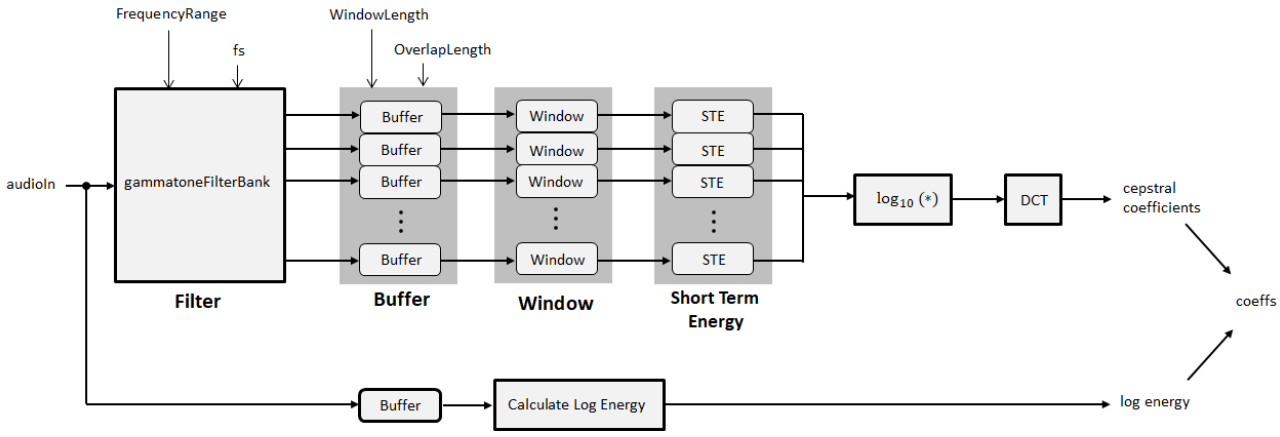


FIGURE 2.5: GFCC Extraction

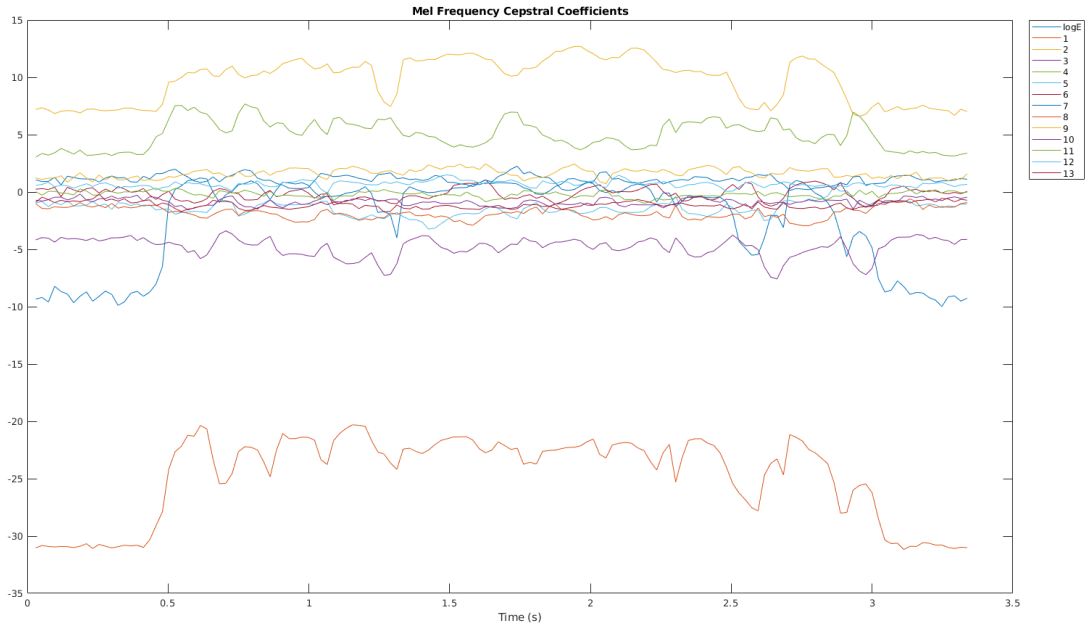


FIGURE 2.6: Typical MFCC plot with log enery and 13 coefficients

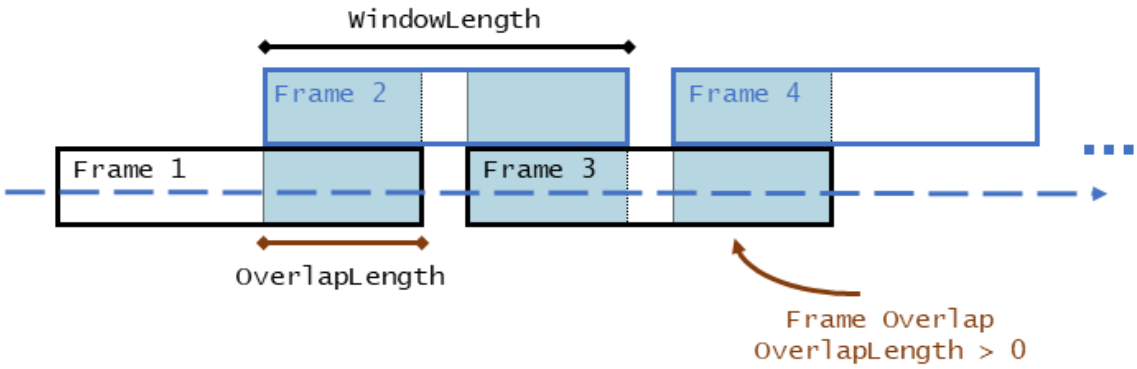


FIGURE 2.7: Audio Feature Extraction using sub-band audio frames

overlap of 60 samples in adjacent frames is considered (See figure:2.7 [15]). Using this methodology, the mel frequency cepstral coefficients, log energy values, cepstral delta, and the cepstral delta-delta values for each segment is calculated. In all 13 mel-frequency cepstral coefficients were considered (See figure: 2.6)

2.3.4 Pitch

Pitch estimates fundamental frequency of audio signal (See figure: 2.8). The pitch values are also estimated for an audio at sub-band levels for frames of length 30ms. For an audio sampled at 8KHz rate, this imply 240 samples per frame. To reduce the effect of windowing, an overlap of 60 samples in adjacent frames is considered (See figure:2.7[15]).

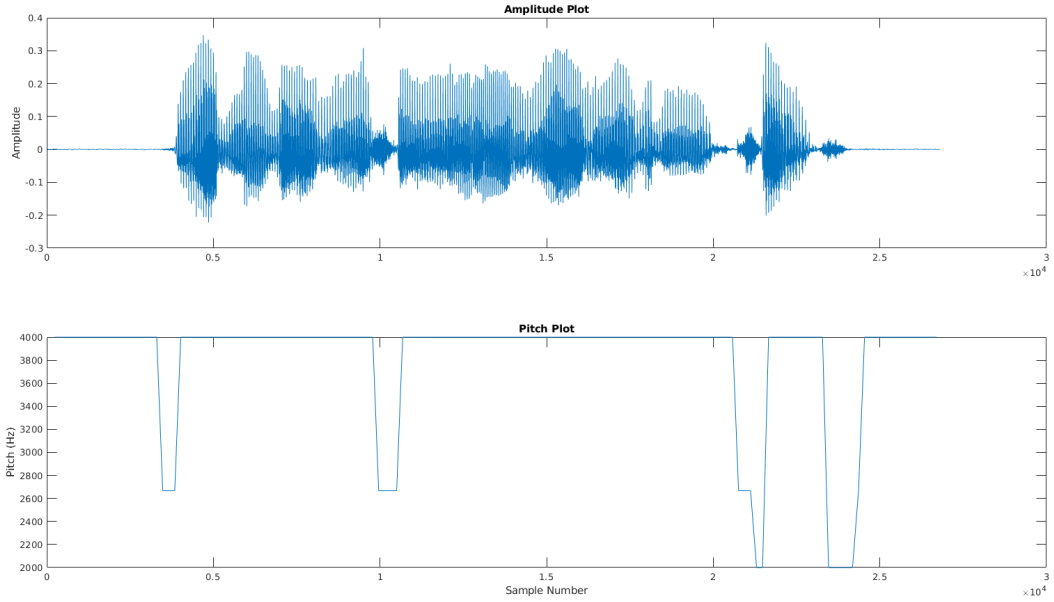


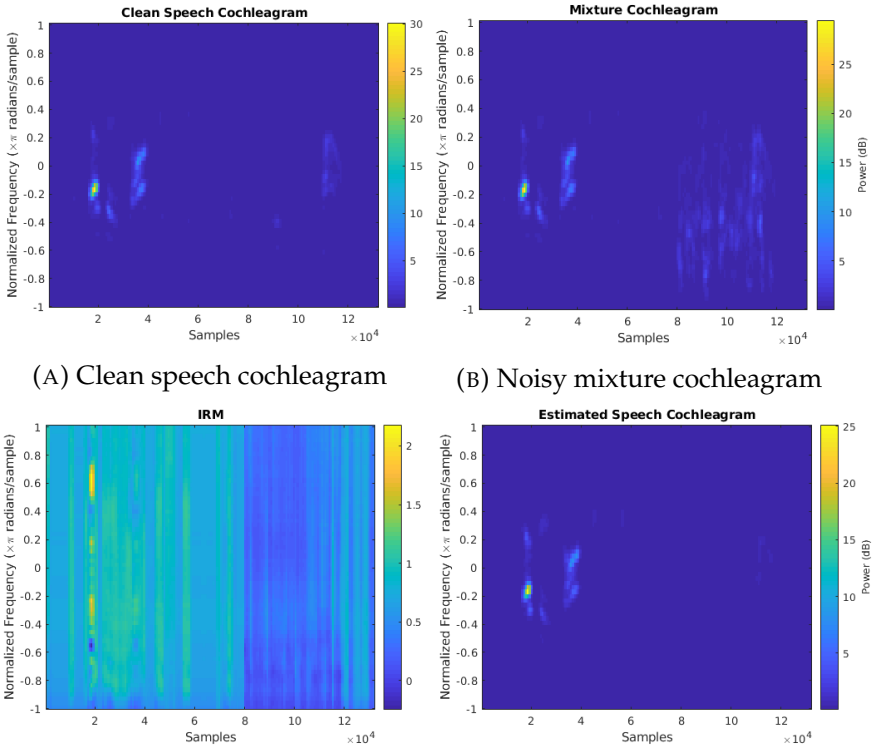
FIGURE 2.8: Pitch plot for an audio waveform

2.3.5 Cochleagram

This is a representation of a time-domain audio waveform in a t-f domain but is different from the spectrogram (See figure:2.9). It is computed using an array of band pass filters that each model the frequency selectivity and nerve response of a single hair cell. Keeping this in mind, a 64 channel Gammatone filter bank as an array of band pass filters is used to estimate cochleagram. Cochleagram evaluation involves using ERB (Equivalent Rectangular Bandwidth) as a psychoacoustic measure for approximating the frequency-dependent bandwidth of the filters in human hearing. The bandwidth of the rectangular bandpass filter is chosen such that it has the same peak and passes the same amount of power for an input of white noise.

2.4 Training Targets

In supervised speech enhancement, defining a proper training target is important for learning and generalization. There are mainly two groups of training targets, i.e., masking-based targets and mapping-based targets. Masking-based targets describe the t-f relationships of clean speech to background interference, while mapping-based targets correspond to the spectral representations of clean speech. We focused on masking based training targets with a particular emphasis on usage of an **IRM** as a training target.



(A) Clean speech cochleagram (B) Noisy mixture cochleagram

(C) IRM - by learning model (D) Estimated speech cochleagram

FIGURE 2.9: Different stages of cochleagram during speech enhancement

2.4.1 IRM

IRM (Ideal Ratio Mask) is a soft mask [16] which can be thresholded into a binary mask based on some local thresholding criterion. IRM is defined as:

$$IRM = \left(\frac{S(t,f)^2}{S(t,f)^2 + N(t,f)^2} \right)^{\beta} \tag{2.1}$$

where

$$S(t,f)^2$$

and

$$N(t,f)^2$$

denote speech energy and noise energy within a T-F unit, respectively. The tunable parameter

$$\beta$$

scales the mask, and is commonly chosen to 0.5. With the square root the IRM preserves the speech energy with each T-F unit, under the assumption that *S(t,f)* and *N(t,f)* are uncorrelated. Without the root the IRM in 2.1 is similar to the classical Wiener filter, which is the optimal estimator of target speech in the power spectrum. An example of the IRM is shown in 2.3c and 2.9c.

Chapter 3

Deep Learning Models

In this project, deep learning models based on following deep neural networks were considered:

- **Feedforward Deep Neural Network**
- **Deep Convolutional Neural Network**

3.1 Baseline DNN model

A baseline deep neural network model was considered with the following architecture (See figure: 3.1):

- **Image Input Layer**:
  This layer accepts the feature data as a 2-D or 3-D input. The number of neurons is this layer depends on the dimension size of the features. For the different combination of features on which this DNN was trained, the minimum number of input neurons were 28 for only GFCC and MFCC features. The maximum number of neurons were 289 for a combination of Spectrogram, GFCC,GFCC Delta ,GFCC double delta, MFCC, MFCC Delta, MFCC double delta and Pitch features.
- **Hidden Layers**:
  Four hidden layers were used with the following properties:
  - **Fully Connected Layer**:
    This layer connect every neuron in one layer to every neuron in another layer [17].
  - **Batch Normalization**:
    Since the activations are constantly changing during training for the intermediate layers, this slows down the training process because each layer must learn to adapt themselves to a new distribution in every training step. This problem is known as "*internal covariate shift*". Batch normalization is a method used to normalize the inputs of each layer, in order to fight the internal covariate shift problem [18]. During training time, a batch normalization layer does the following:
    1. Calculate the mean and variance of the layers input.
    2. Normalize the layer inputs using the previously calculated batch statistics.
    3. Scale and shift in order to obtain the output of the layer.

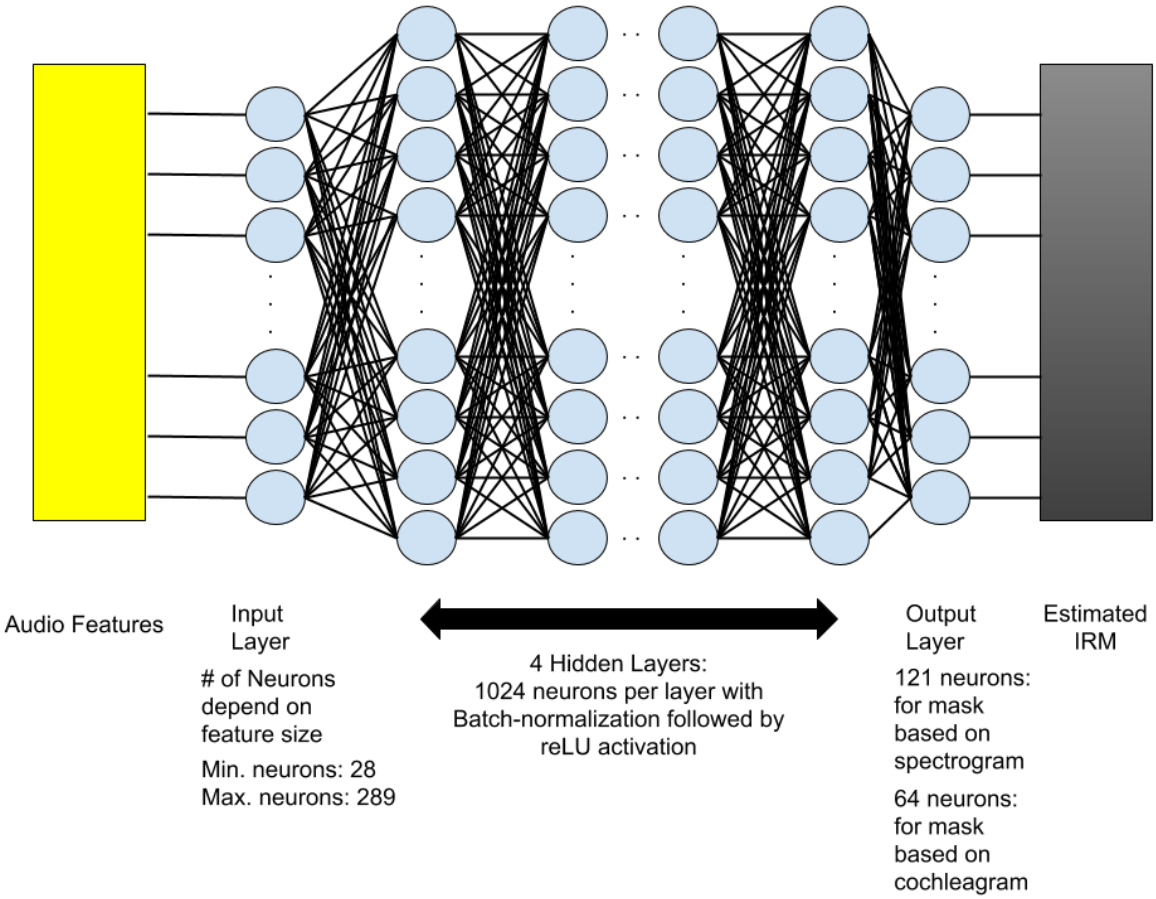


FIGURE 3.1: A Baseline feedforward deep neural network with 5 layers

– **reLU activation**:
The rectified linear unit activation function is a piecewise linear function that will output the input directly if is positive, otherwise, it will output zero [19]. This thus helps in overcoming "*vanishing gradient problem*" (See 3.3) often encountered with traditional tanh or sigmoid activations. reLU activation applys the function

$$f(x) = max(0, x)$$

where x is the output calculated post weight multiplication and bias addition at a given layer.

- **Output Layer**:
Output layer is a fully connected layer with regression operation. The number of neurons in this layer is equal to the dimension of the IRM to estimate. For a spectrogram based IRM, the number of output neurons are 121 and for a cochleagram based IRM, the number of output neurons are 64. Regression operation tries to minimize a cost function to model the feature set with respect to the IRM to estimate. The metric used for optimisation is Root Mean Square Error (RMSE) to minimize the intended cost function for modeling the training data.

3.1.1 Experiments with baseline DNN

Baseline DNN was used to identify the best feature set for speech enhancement with an emphasis on finding simpler features with improvement in the intelligibility and quality with reduced computing resource consumption. Experiments were divided into two categories based on training targets, viz "IRM based on spectrograms" and "IRM based on cochleagram". Baseline DNN was trained for noisy mixtures at 0 SNR and -2 SNR respectively.

1. **IRM based on spectrograms**:

- **Intelligibility**:
STOI performance for different combination of audio features is listed in the table 3.1. Following inferences can be made:

(a) **Comparison between worst performing and best performing features**:
As evident from the table 3.1, the gain in intelligibility from the worst performing and the best performing feature set is "**2.5%**" for SNR 0 and "**1.3%**" for SNR -2.

(b) **Comparison of intelligibility with noisy audios**:
On comparing with the intelligibilty scores of the original noisy mixture for the best performing feature set (See table: 3.6) the gain in intelligibility scores for SNR 0 is "**10.9%**" and for SNR -2 is "**14.9%**".

- **Quality**:
PESQ performance for different combination of audio features is listed in

Features	STOI	
	SNR 0	SNR -2
Spectrogram	0.81	0.76
Spectrogram, MFCC, MFCC Delta, MFCC delta delta	0.80	0.77
Spectrogram, Pitch	0.79	0.76
Spectrogram, GFCC, GFCC Delta, GFCC delta delta	0.80	0.77
Spectrogram, GFCC, GFCC Delta, GFCC delta delta, MFCC, MFCC Delta, MFCC delta delta	0.80	0.77
GFCC, MFCC	0.79	0.76
Spectrogram, GFCC, MFCC	0.80	0.77
GFCC, GFCC Delta, GFCC delta delta	0.81	0.77
GFCC, GFCC Delta, GFCC delta delta, MFCC, MFCC delta, MFCC delta delta	0.81	0.77
Spectrogram, GFCC, GFCC delta, GFCC delta delta, MFCC, MFCC delta, MFCC delta delta, Pitch	0.79	0.76
MFCC, MFCC delta, MFCC delta delta	0.80	0.77

TABLE 3.1: STOI performance: Baseline, IRM based on spectrogram

the table 3.2.Following inferences can be made:

(a) **Comparison of quality gain from worst to best performing feature set**:
The quality gain from the worst performing feature set to the best performing feature set is "**9.9%**" for **PESQMOS** and "**11.5%**" for **MOSLQO** for SNR 0. For SNR -2 the gain in quality is "**6.6%**" for **PESQMOS** and "**8.8%**" for **MOSLQO** respectively.

(b) **Comparison of quality gain from noisy mixtures**:
It's also important to compare the gain in quality in estimated speech signal as compared to the quality scores of the noisy mixture which has been tabulated in 3.3 for the best performing feature set, for SNR 0 and -2 respectively. Hence, comparing both tables gives us a gain in quality of "**10.5%**" for **PESQMOS** and "**10.9%**" for **MOSLQO** at SNR 0 and the gain in quality of "**9.8%**" for **PESQMOS** and "**11.3%**" for **MOSLQO** at SNR -2 respectively.

2. **IRM based on cochleagrams**:

- **Intelligibility**:
STOI performance for different combination of audio features is listed in the table 3.4. Following inferences can be made.

(a) **Comparison with the IRM based on Spectrogram**:
As evident from the table 3.4, the gain in intelligibility score for the best performing feature sets is "**1.2%**" when compared to the best performing feature set for the IRM based on spectrogram for SNR 0 and for SNR -2 the intelligibility scores are the same for both the cases.

(b) **Comparison from the worst to best performing feature set**:
The gain in intelligibility score from worst performing feature set to

Features	PESQ			
	SNR 0		SNR -2	
	PESQMOS	MOSLQO	PESQMOS	MOSLQO
Spectrogram	2.11	1.82	2.10	1.81
Spectrogram,MFCC,MFCC delta,MFCC delta delta	2.31	2.03	2.21	1.93
Spectrogram,Pitch	2.23	1.95	2.18	1.93
Spectrogram,GFCC,GFCC delta,GFCC delta delta	2.29	1.99	2.24	1.97
Spectrogram,GFCC,GFCC delta,GFCC delta delta MFCC,MFCC delta,MFCC delta delta	2.30	2.02	2.19	1.92
GFCC,MFCC	2.26	1.96	2.16	1.88
Spectrogram,GFCC,MFCC	2.29	2.00	2.22	1.95
GFCC,GFCC delta,GFCC delta delta	2.32	2.03	2.20	1.92
GFCC,GFCC delta,GFCC delta delta MFCC,MFCC delta,MFCC delta delta	2.28	1.99	2.19	1.90
Spectrogram,GFCC,GFCC delta,GFCC delta delta MFCC,MFCC delta,MFCC delta delta,Pitch	2.19	1.90	2.16	1.87
MFCC,MFCC delta,MFCC delta delta	2.26	1.98	2.13	1.85

TABLE 3.2: PESQ performance: Baseline, IRM based on spectrogram

Features	PESQ			
	SNR 0		SNR -2	
Noise mixed speech samples	2.10	1.83	2.04	1.73

TABLE 3.3: PESQ values for noisy mixture

the best performing feature set is by "**6.5%**" for SNR 0 and "**6.9%**" for SNR -2.

(c) **Gain in intelligibility from noisy mixtures**:
For the best perfroming feature set, when compared to the original noisy mixture's intelligibility (see table: 3.6), the gain is "**12.3%**" for SNR 0 and for SNR -2, the gain in intelligibility is "**14.9%**".

Hence, it highlights a better performance than the IRM based on spectrograms for the intelligibility gains.

- **Quality**:
PESQ performance for different combination of audio features is listed in the table 3.5. Following observations can be made:

(a) **Comparison with IRM based on spectrogram**:
As it's evident from the table 3.5, the quality scores are better as compared to IRM based on spectrogram for the best performing feature sets in the both cases by around 2.1% for 0 SNR and by around 3.1% for -2 SNR considering PESQMOS as quality metric for the best performing feature set.

(b) **Comparison of quality gain from worst to best performing feature set**:
Quality gain from the worst performing feature set to the best performing feature set is by "**9.7%**" for **PESQMOS** and "**11.9%**" for **MOSLQO** at SNR 0. At SNR -2 the gain in quality is "**13.8%**" for **PESQMOS** and "**14.3%**" for **MOSLQO**.

(c) **Comparison with the quality of noisy mixtures**:
On comparing for quality gain from the noisy mixtures for the best performing feature set as tabulated in the table 3.3, there is a gain of "**12.8%**" for **PESQMOS** and "**13.1%**" for **MOSLQO** at SNR 0 and the gain in quality of "**13.2%**" for **PESQMOS** and "**15.6%**" for **MOSLQO** at SNR -2 respectively.

This highlights the importance of "**Cochleagram**" as training data for low SNR scenarios and also highlight the overall improvement in quality when using cochleagram based masks as compared to spectrogram based masks during training.

Features	**STOI**	
	SNR 0	SNR -2
Cochleagram	0.82	0.77
GFCC,GFCC delta,GFCC delta delta	0.78	0.73
GFCC,GFCC delta,GFCC delta delta MFCC,MFCC delta,MFCC delta delta	0.77	0.74
MFCC,MFCC delta,MFCC delta delta	0.78	0.75
GFCC,MFCC	0.78	0.74

TABLE 3.4: STOI performance: Baseline, IRM based on cochleagram

Features	PESQ			
	SNR 0		SNR -2	
	PESQMOS	MOSLQO	PESQMOS	MOSLQO
Cochleagram	2.37	2.07	2.31	2.00
GFCC,GFCC delta,GFCC delta delta	2.23	1.92	2.05	1.75
GFCC,GFCC delta,GFCC delta delta MFCC,MFCC delta,MFCC delta delta	2.21	1.91	2.09	1.80
MFCC,MFCC delta,MFCC delta delta	2.16	1.85	2.09	1.82
GFCC,MFCC	2.27	1.96	2.06	1.77

TABLE 3.5: PESQ performance: Baseline, IRM based on cochleagram

Features	STOI	
	SNR 0	SNR -2
Noise Mixed Speech Samples	0.73	0.67

TABLE 3.6: STOI values for Noise Mixed speech samples

3.2 DNN model based on biased sigmoid activation

Another deep neural network model was considered with the following architecture (See figure: 3.3):

- **Image Input Layer**:
  This layer accepts the feature data as a 2-D or 3-D input. The number of neurons is this layer depends on the dimension size of the features just as with the baseline model.

- **Hidden Layers**:
  Four hidden layers were used with the following properties:

  - **Fully Connected Layer**:
    This layer connect every neuron in one layer to every neuron in another layer.
  - **Biased Sigmoid Layer**:
    This layer is an alteration on a sigmoid layer which uses a sigmoid activation function (See figure: 3.2) to overcome "*vanishing gradient problem*". The number of neurons in these layers are twice the number of neurons in the input layer. Sigmoid function, squishes a large input space into a small input space between 0 and 1. Therefore, a large change in the input of the sigmoid function will cause a small change in the output. Hence, the derivative becomes small. This causes the neural network to have a reduced learning with the gradients not able to reach the optimization points over time. The biased sigmoid activation layer overcomes this problem by enforcing the sigmoid function to operate on it's linear part where the derivative of the function is significantly better than the case when the sigmoid function operates closer to the region of 0 and 1. Hence, the biased sigmoid layer is able to overcome the "*Vanishing Gradient Problem*". The purpose of using this layer is to compare it's performance with the reLU activation used in the baseline DNN as discussed in the previous section.

– **Batch Normalization**:
As discussed in the previous section, this layer helps in overcoming"*internal covariate shift*". It does so by normalizing the inputs of each layer.

– **Dropout Layer**:
This layer works on a tunable parameter called the "*Dropout Probability*". During the training phase, the proportion of neurons as defined by the parameter "*Dropout Probability*" on a particular layer are deactivated. This improve generalization because it forces a layer to learn with different neurons the same "concept". Hence, this is used to prevent overfitting of the training data, where a learning machine has an ability to model the training data but performs poorly on the testing data. During the prediction phase however, the dropout is deactivated.

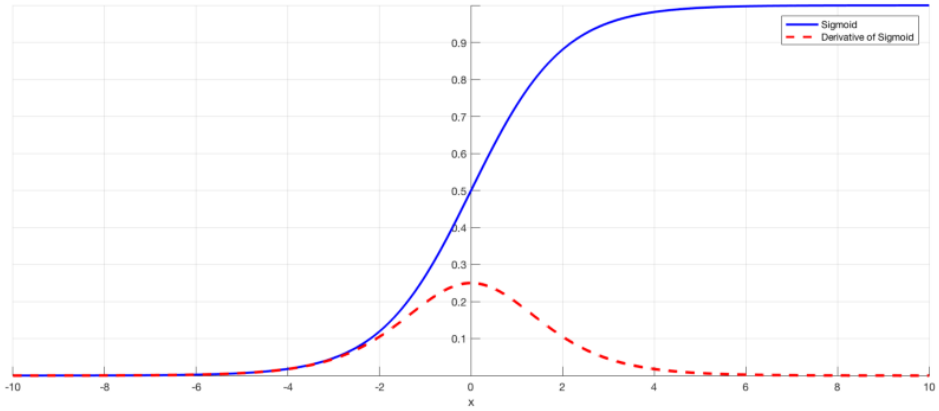


FIGURE 3.2: Sigmoid activation function.

- **Output Layer**:
Output layer is a fully connected layer with biased sigmoid activation and regression operation. The number of neurons in this layer is equal to the dimension of the IRM to estimate. For a spectrogram based IRM, the number of output neurons are 121 and for a cochleagram based IRM, the number of output neurons are 64. Biased sigmoid layer tries to prevent "*Vanishing Gradient Problem*" followed by a regression operation, which tries to minimize a cost function to model the feature set with respect to the IRM to estimate by using RMSE as an optimization metric.

3.2.1 Experiments with the DNN using biased activation

Using the results from the baseline DNN, experiments were conducted using the best performing feature sets with the baseline DNN. Similar to the baseline DNN, experiments were divided into two categories based on training targets, viz "IRM based on spectrograms" and "IRM based on cochleagram". Training was done for noisy mixtures at 0 SNR and -2 SNR respectively.

1. **IRM based on Spectrogram**

   - **Intelligibility**:
   STOI performance for different combination of audio features is listed in

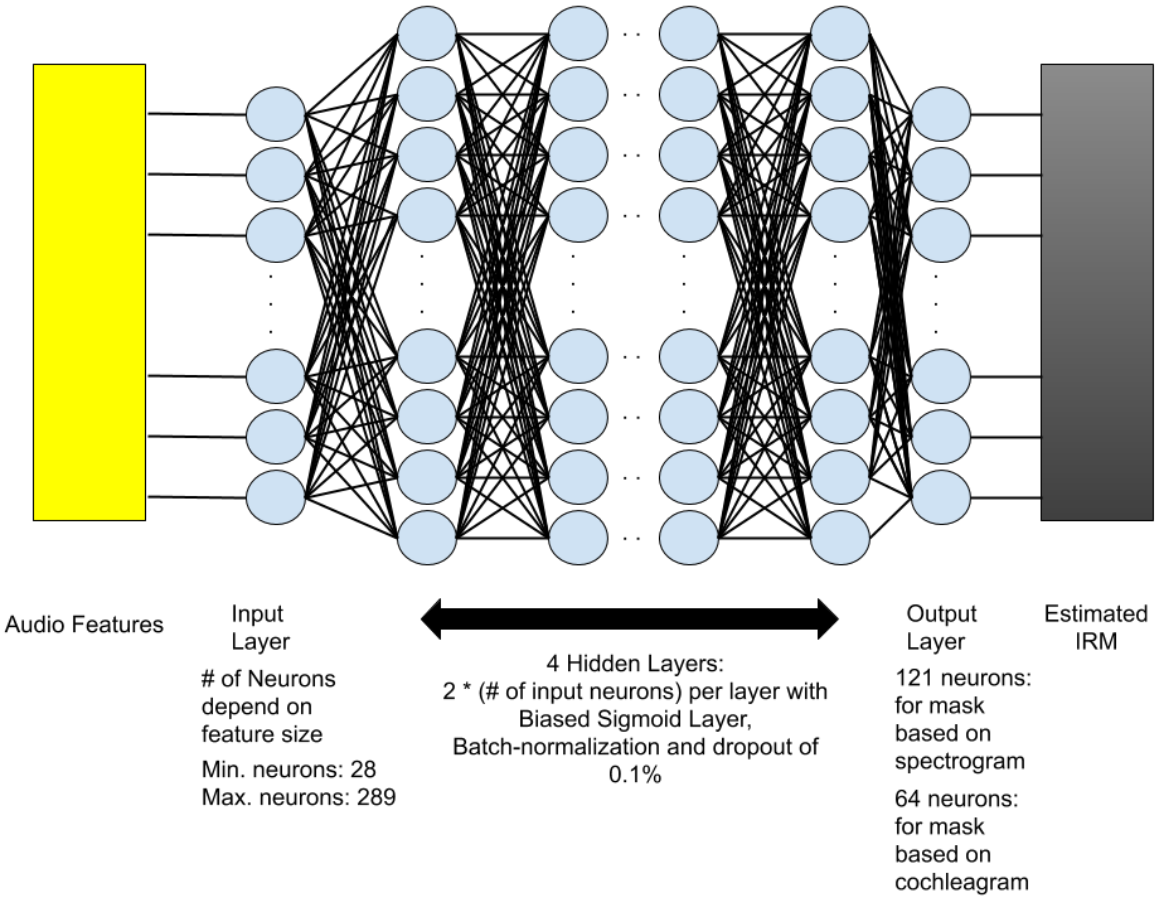


FIGURE 3.3: A feedforward deep neural network with 5 layers and biased sigmoid activation

the table 3.7. Following observations can be made:

(a) **Comparison with the baseline DNN**:
As it's evident from the table 3.7, there is no significant improvement in the intelligiblity scores as compared to the baseline DNN's performance tabulated in the table 3.4. Infact for the best performning feature set, the intelligibilty score is reduced by "**1.25%**" for SNR 0. Although the intelligibility for SNR -2 as compared to the baseline DNN is the same for the same feature set.

(b) **Comparison of intelligibility gain from worst to the best performing features**:
The improvement in intelligibility for the feature sets used for this experiment from the worst performing to the best performing set is by "**2.5%**" for SNR 0 and for SNR -2, the improvement is "**4%**".

(c) **Comparison with the noisy mixtures**: The intelligibility gains as compared to the noisy mixtures (see table: 3.6) is "**9.6%**" for SNR 0 and for SNR -2 the gain is of "**14.9%**".

The best performing feature set in this case was still found out to be "**GFCC, GFCC delta, GFCC delta delta**" for both SNR 0 and SNR -2 which is similar to the performance of this feature set with the baseline DNN for SNR 0 whereas in the case of the baseline DNN for -2 SNR, the best feature set was "**Spectrogram, GFCC, GFCC delta, GFCC delta delta**". This reflects a strong correlation of GFCC related feature set with the IRM based on spectrogram. However, overall the intelligibilty performance shows no significant improvement as compared to the baseline DNN for this deep learning model.

Features	**STOI**	
	SNR 0	SNR -2
Spectrogram	0.79	0.76
Spectrogram,GFCC,GFCC delta,GFCC delta delta	0.80	0.76
Spectrogram,GFCC,MFCC	0.78	0.77
GFCC,GFCC delta,GFCC delta delta	0.80	0.77
GFCC,GFCC delta,GFCC delta delta MFCC,MFCC delta,MFCC delta delta	0.79	0.74

TABLE 3.7: STOI performance: DNN2, IRM based on spectrogram

- **Quality**:
PESQ performance for different combination of audio features is listed in the table 3.8. Following inferences can be made:

(a) **Comparison with the baseline DNN**:
From the table 3.8, and comparing the PESQ scores with the baseline DNN documented in the table 3.2, the scores show deterioration of the quality. The best performing feature set "*Spectrogram,GFCC,GFCC delta,GFCC delta delta*", shows detereoration of "**0.8%**" for **PESQMOS** score and detereoration of "**0.9%**" for the **MOSLQO** score when compared to the performance of the best performing feature set found for the baseline DNN at SNR 0. For SNR -2, the best performing feature

set "**GFCC, GFCC delta, GFCC delta delta**" showed the detereoration of "**1.8%**" for the **PESQMOS** score, and the **MOSLQO** scores were the same at "1.97".

(b) **Comparison of quality gain from worst to the best performing feature**:
For this experiment the gain of quality from the worst performing feature set to the best performing feature set was of "**8.4%**" for **PESQMOS** and "**5.2%**" for **MOSLQO** at SNR 0. At SNR -2, the gain in quality from the worst to the best performing feature set was of "**10%**" for **PESQMOS** and "**8.2%**" for **MOSLQO**.

(c) **Comparison with the noisy mixture**:
On comparing the quality gain for the best performing feature set in this experiment from the noisy mixture as documented in the table 3.8 the gain was "**9.5%**" for **PESQMOS** and "**9.8%**" for **MOSLQO** at SNR 0. At SNR -2, the gain was "**7.8%**" for **PESQMOS** and "**11.3%**" for **MOSLQO**.

Overall, it can be said that there is no significant improvement when comapred to the quality gain performance of the baseline DNN for this deep learning model.

Features	**PESQ**			
	SNR 0		SNR -2	
	PESQMOS	MOSLQO	PESQMOS	MOSLQO
Spectrogram	2.21	1.91	2.17	1.90
Spectrogram,GFCC,GFCC delta,GFCC delta delta	2.30	2.01	2.12	1.82
Spectrogram,GFCC,MFCC	2.22	1.98	2.00	1.93
GFCC,GFCC delta,GFCC delta delta	2.28	1.99	2.20	1.97
GFCC,GFCC delta,GFCC delta delta MFCC,MFCC delta,MFCC delta delta	2.12	1.97	2.00	1.89

TABLE 3.8: PESQ performance: DNN2, IRM based on spectrogram

2. **IRM based on Cochleagram**

- **Intelligibility** STOI performance for different combination of audio features is listed in the table 3.9. Following observations can be made:

(a) **Comparison with the baseline DNN**:
From the table 3.9, shows a detereoration in the intelligibility score as comapred to the baseline DNN's performance by "**2.5%**" for SNR 0 and for SNR -2, by "**1.3%**" for the best performing feature set.

(b) **Comparison from the worst and the best performing feature set**:
Since the intelligibility performance was poor with the feature sets in the previous experiment with IRM based on spectrogram, this experiment was confined to the best performing set of "*GFCC, GFCC delta, GFCC delta delta*" and "*Cochleagram*". Within this experiment, the intelligibility gain from worst performing feature set to the best performing feature set is "**3.9%**" for SNR 0 and for SNR -2, the gain was "**5.5%**".

(c) **Comparison with the noisy mixtures**:
On comparing with the intelligibility of the noisy mixtures (see table: 3.6), the gain was found out to be "**9.6%**" at 0 SNR and at "**13.4%**" -2 SNR for "*Cochleagram*" as the better performing feature set in the previous case.

This shows an improvement in intelligibility scores when using IRM based on cochleagram as training target compared to the IRM based on spectrogram which is as per the trend with the baseline DNN as well. However, compared to intelligibility gains provided by the baseline DNN, this deep learning model shows no significant improvement.

Features	**STOI**	
	SNR 0	SNR -2
Cochleagram	0.80	0.76
GFCC, GFCC delta, GFCC delta delta	0.77	0.72

TABLE 3.9: STOI performance: DNN2, IRM based on cochleagram

- **Quality** PESQ performance for the different combination of audio features is listed in the table 3.10. Following inferences can be made:

(a) **Comparison with the baseline DNN**:
On comparing the tabulated data from the table 3.10, with the baseline DNN's performance in the table 3.5, for the best performing feature set, there is detereoration in quality by "**1.7%**" for **PESQMOS** and "**0.9%**" for **MOSLQO** for SNR 0. For SNR -2, the detereoration is by "**1.7%**" for **PESQMOS** and "**1%**" for **MOSLQO**.

(b) **Comparison from the worst to best performing feature set**:
Within this experiment the gain in quality from the worst perfroming feature set to the best performing feature set was found out to be "**8.1%**" for **PESQMOS** and "**9.9%**" for **MOSLQO** for SNR 0. For SNR -2, the detereoration is by "**11.8%**" for **PESQMOS** and "**13.1%**" for **MOSLQO**.

(c) **Comparison with the noisy mixture**:
On comapring with the quality of noisy mixtures as tabulated in the table 3.3, the gain in quality was found out to be "**11.2%**" for **PESQMOS** and "**12.9%**" for **MOSLQO** for SNR 0. For SNR -2, the detereoration is by "**13.5%**" for **PESQMOS** and "**13.8%**" for **MOSLQO** respectively.

Overall, this too shows no significant improvement in quality gains as comaperd with the gains provided by the baseline DNN.

Features	**PESQ**			
	SNR 0		SNR -2	
	PESQMOS	MOSLQO	PESQMOS	MOSLQO
Cochleagram	2.39	2.10	2.27	1.98
GFCC, GFCC delta, GFCC delta delta	2.21	1.91	2.03	1.75

TABLE 3.10: PESQ performance: DNN2,IRM based on cochleagram

3.3 Deep Learning Model based on CNN

A deep learning model based on Convolutional Neural Networks (See figure: 3.4) places less demand on the features as they have an ability to find patterns in the training data by themselves. Exploiting this nature, the experiments with this learning model were confined to the two prominent visual representations of the audio signals namely, Spectrograms and Cochleagrams. The architecture of the considered deep CNN was as following:

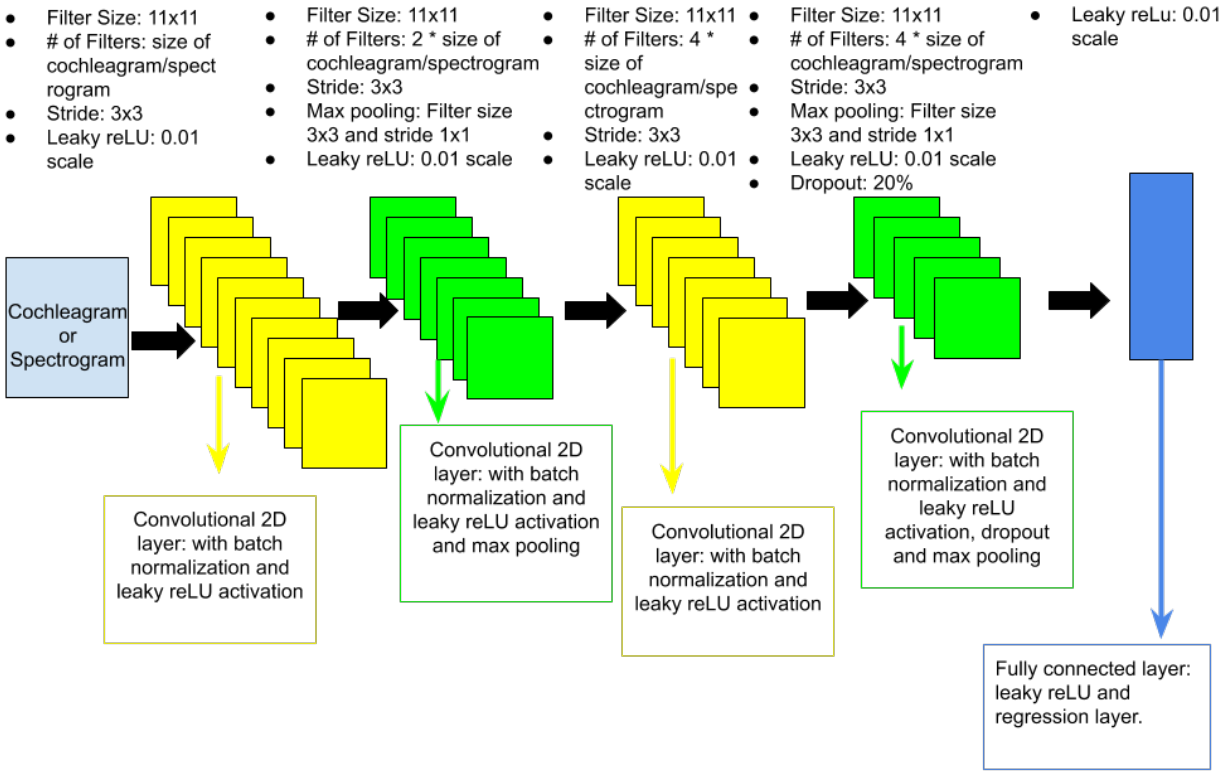


FIGURE 3.4: A CNN with 4 2-D Convolutional Layers and 2 downsampling 2-D Maxpooling Layers

- **Image Input Layer**:
  As described in the previous sections, this layer accepts the feature data as a 2-D or 3-D input. The number of neurons is this layer depends on the dimension size of the features. For cochleagram based training data, the number of input neurons are 64 and for spectrogram based training data, the number of input neurons are 121.

- **Hidden Layers**:
  Hidden layers in considered deep CNN were combination of **four** convolution2dLayer(s) with a downsampling by maxPooling2dLayer(s) after second and fourth convolution2dLayer respectively. The hidden layers had the following properties:

- **Convolution2dLayer**:
  This is a 2-D convolutional layer which applies sliding convolutional filters to the input. The layer convolves the input by moving the filters along the input vertically and horizontally and computing the dot product of the weights and the input, and then adding a bias term. The convolutional **filter's kernel size** was kept at **11x11** and the **number of convolutional filters** were kept the **same as feature size i.e 64/121** for the **first** convolution2dLayer, **2x(feature size i.e 64/121)** for the **second** convolution2dLayer and **4x(feature size i.e 64/121)** for the **remaining two** convolution2dLayer(s).
- **Batch Normalization**: As discussed in the previous sections, this layer helps in overcoming "internal covariate shift" problem. It does so by normalizing the inputs of each layer.
- **Leaky reLU activation**:
  Using reLU activation often encounters a "*Dying ReLU problem*" i.e. when inputs approach zero, or are negative, the gradient of the function becomes zero, the network cannot learn these inputs. Leaky reLU prevents the dying ReLU problem by making a slight variation on ReLU by having a small positive slope in the negative area (See figure: 3.5), so it does enable the network to learn, even for negative input values. The leaky reLU uses the following activation function to achieve this using the scale value of 0.01 as was considered in our implementation:

$$f(y) = \begin{cases} y & \text{if } y > 0 \\ 0.01y & \text{otherwise} \end{cases}$$

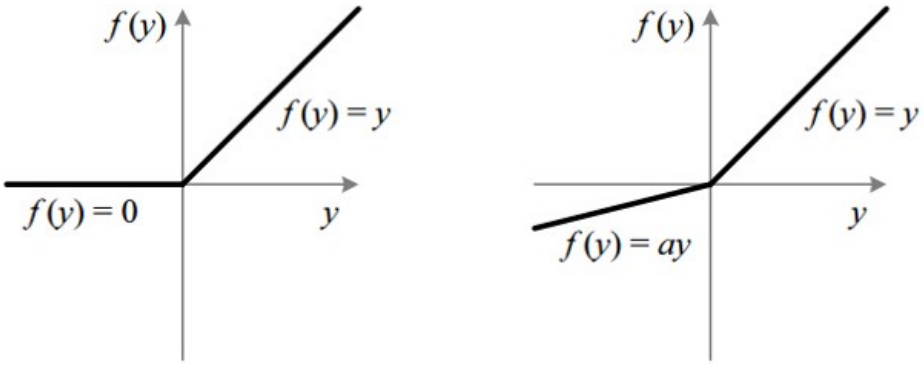


FIGURE 3.5: Leaky reLU as compared to reLU

- **MaxPooling2dLayer**:
  This layer performs the down-sampling by dividing the input into rectangular pooling regions, and computing the maximum of each region. In our implementation, the max pooling filter of size **3x3** was considered.

- **Output Layer**:
  Output layer is a fully connected layer with leaky reLU activation and regression operation. For a spectrogram based IRM, the number of output neurons are 121 and for a cochleagram based IRM, the number of output neurons are

64. Regression operation tries to minimize a cost function to model the spectrogram/cochleagram with respect to the IRM to estimate. The metric used for optimisation is RMSE to minimize the intended cost function for modeling the training data.

3.3.1 Experiments with the deep CNN

The CNN was trained using spectrogram and cochleagram as training data. The intended training target for spectrogram was IRM based on spectrogram and for cochleagram was IRM based on cochleagram respectively. The performance analysis are as below:

- **Intelligibility**:
  The STOI values for 0 and -2 SNR are documented in the table 3.11. As per the information from this table following observations can be made:

  1. **Comparison with the baseline DNN**:
     For the best performing feature set, the gain in intelligibilty has been observed for CNN based learning model to be of **4.8%** and **3.9%** for 0 and -2 SNR respectively.
  2. **Comparison between worst and best performing training data**:
     The gain in intelligibility from the worst to the best performing training data is **8.8%** and **5.3%** for 0 and -2 SNR respectively. Since, only spectrogram and cochleagram was used as the training data, this is a comparison of intelligibility gain from spectrogram to cochleagram as training data.
  3. **Intelligibility gain from the noisy mixture audios**:
     The gain in intelligibility when considering the original noisy mixture (See table: 3.6) for the best performing training data was found out to be **17.8%** and **19.4%** for 0 and -2 SNR respectively.

  This points out that for cochleagram based training data and IRM based on the cochleagram the intelligibility gain is the best when compared to the previous learning models.

Features	**STOI**	
	SNR 0	SNR -2
Cochleagram	0.86	0.80
Spectrogram	0.79	0.76

TABLE 3.11: STOI performance: CNN

- **Quality**:
  The PESQ values for 0 and -2 SNR are documented in the table 3.12. As per the information from this table following observations can be made:

  1. **Comparison with the baseline DNN**:
     On comparing the tabulated data from the table 3.12, with the baseline DNN's performance in the table 3.5, for the best performing feature set, there is gain in quality by "**2.1%**" for **PESQMOS** and "**6.7%**" for **MOSLQO** at SNR 0. For SNR -2, the gain in quality is by "**2.2%**" for **PESQMOS** and "**8%**" for **MOSLQO**.

2. **Comparison between worst and best performing training data**:
   The gain in quality from the worst to the best performing training data is "**9%**" for **PESQMOS** and "**14.5%**" for **MOSLQO** at SNR 0. For SNR -2, the gain in quality is by "**12.3%**" for **PESQMOS** and "**16.7%**" for **MOSLQO**.
3. **Quality gain from the noisy mixture audios**: On comapring with the quality of noisy mixtures as tabulated in the table 3.3, the gain in quality is found out to be "**15.2%**" for **PESQMOS** and "**20.7%**" for **MOSLQO** for SNR 0. For SNR -2, the gain in quality is by "**15.6%**" for **PESQMOS** and "**22.4%**" for **MOSLQO** respectively.
   This again points out that for cochleagram based training data and IRM based on the cochleagram the quality gain is the best when compared to the previous learning models.

Features	**PESQ**			
	SNR 0		SNR -2	
	PESQMOS	MOSLQO	PESQMOS	MOSLQO
Cochleagram	2.50	2.21	2.38	2.16
Spectrogram	2.22	1.93	2.10	1.85

TABLE 3.12: PESQ performance: CNN

Chapter 4

Performance Analysis and Conclusions

4.1 Intelligibility Analysis

On comparing the best performing training data for the three deep learning models as discussed in the chapter 3, the gain in the intelligibility can be tabulated as below:

Deep Learning Model	**STOI Gain at 0 SNR**	**STOI Gain at -2 SNR**
Baseline DNN	12.3%	14.9%
DNN with biased sigmoid activation	9.6%	14.9%
CNN with leaky reLU	17.8%	19.4%

TABLE 4.1: Intelligibility Gains

Hence, from 4.1 and observing the intelligibility scores of original noisy mixtures from the table 3.6, we can make following analysis:

- STOI for original noisy mixture at 0 SNR is 0.73. This is already a decent score (on the STOI metric scale of 0-1) and hence, the subsequent gain in intelligibility at 0 SNR post speech enhancement is lower than the STOI gain at -2 SNR for which the original noisy mixture has a STOI score of 0.67.
- There is a clear trend of improvement of STOI gain from the noisy audio mixtures at 0 SNR to the noisy audio mixtures at -2 SNR. This affirms the ability of IRM in preserving the clean speech energies which is helpful for speech enhancement particularly in low SNR scenarios.
- Second DNN with the biased sigmoid activation provides less gain in intelligibility scores for 0 SNR and the same score of intelligibility for the -2 SNR when compared to the baseline DNN. It's intelligibility gain when compare to the CNN based model is however poor. Using baseline DNN as a reference, for the dataset considered, the prima facie trend show slightly poor performance of the biased sigmoid layer comapred to the reLU activation used in the baseline. But, to make a conclusive remark training on a larger dataset would be required.
- CNN with leaky reLU shows the best intelligibility gains among the three considered deep learning models and also places the least demand on the feature extraction as only spectrogram/cochleagram were used in the training phase.

- For DNN based learning models and the spectrogram based training target, the better performing feature set was always found to include the gammatone cepstral features i.e. the set "*GFCC, GFCC delta, GFCC delta delta*". However, for the training target based on cochleagram, the best performing feature set was always found to include the feature set "*Cochleagram*".

4.2 Quality Analysis

On comparing the best performing training data for the three deep learning models as discussed in the previous chapter 3, the gain in the quality scores considering MOSLQO has been tabulated in the table 4.2. MOSLQO is chosen over PESQMOS as it defines perceptual quality gain with respect to the human listening objectivity.

Deep Learning Model	**Quality Gain at 0 SNR**	**Quality Gain at -2 SNR**
Baseline DNN	13.1%	15.6%
DNN with biased sigmoid activation	12.9%	13.8%
CNN with leaky reLU	20.7%	22.4%

TABLE 4.2: Quality Gains

Hence from the tables 4.2 and 3.3, the following analysis can be made for quality as a performance metric:

- MOSLQO scores for original noisy mixture at 0 SNR is 1.83. This is qualitatively at the lower scale of 1-5 for MOSLQO. The subsequent gain in quality metric at 0 SNR shows an improvement of 15.6% on average. Likewise, quality scores at -2 SNR for original noisy mixtures is 1.73, but the quality gain is higher as compared to the 0 SNR case with an average gain of 17.2%. This points out, and as also observed from all three learning models, the speech enhancement has more scope of audio quality improvement at the lower SNR levels when using IRM as the training target.

- DNN with biased sigmoid activation provides the least gain in the quality among the three deep learning models considered and hence, strengthen the argument that the reLU activation considered in the baseline DNN performs better than the biased sigmoid activation. However, to conclusively make this statement and as per the analysis from the intelligibility gains, training of both models on a larger dataset would be required.

- CNN with leaky reLU shows the best quality gains among the three considered deep learning models along with the intelligibility as previously observed and also places the least demand on the feature extraction. This reaffirms the ability of this category of deep learning machines in better speech enhancement as compared to the other models.

4.3 Conclusions

Based on study of the experiments conducted the following conclusions can be made:

- DNN based learning models perform better with cochleagram as training data and IRM based on cochleagram as the training target when compared with audio features and IRM based on spectrogram.
- Using the feature set "GFCC, GFCC delta, GFCC delta delta" has the best performance for 0 SNR with DNN based learning models and IRM based on spectrogram as the training target.
- Using the feature set "GFCC, GFCC delta, GFCC delta delta, Spectrogram" has the best performance for -2 SNR with DNN based learning models and IRM based on spectrogram as the training target.
- This highlights the importance of gammatone based cepstral features in speech enhancement where the learning machine places a significant demand on feature extraction as a part of training data.
- CNN with leaky reLU activation has the best performance in terms of gains in intelligibility and quality scores out of all considered models.
- Performance of CNN also underlines the importance and power of this category of learning machines which have the ability to learn the patterns in the training data by themselves thus placing minimum demand on feature extraction.
- It's also noteworthy that the improvement in gain of both intelligibility and quality post speech enhancement is more in -2 SNR as compared to 0 SNR. Hence, for noisy mixtures with even lower SNR values, the scope of intelligibility and quality gain is established as per the observed trends from the experiments.

4.4 Future Prospects

- **Train on larger dataset**:
  All of the considered models were trained on the TIMIT dataset of 1718 audio samples. Hence, by expanding the dataset all of the models can be trained to generalize and learn "noise" from the noisy mixtures more efficiently thus helping in better speech enhancement and source separation. This can thus help in improvement of intelligibility and quality gains.
  This would also help in conclusively answering the argument regarding the performance of the biased sigmoid layer with respect to the reLU layer.
- **Train on lower SNR values**:
  All of the considered models were trained for noisy mixtures at 0 and -2 SNR respectively. As per the analysis of performanc metrics, the gain in quality and intelligibility was more for noisy mixtures at -2 SNR post enhancement. By testing on noisy mixtures at even lower SNRs, this trend can be verified.
- **Scale to binaural models**:
  Using the Interaural Time Difference (ITD) and Interaural Intensity Difference(IID) as additional features, these models can be trained for binaural mixtures.

- **Testing on more models**:
  This involves exploring unsupervised learning as standalone or in tandem with the supervised learning to check for any improvement in performance metrics.

- **Weight Initialization**:
  In the models implemented, the weights are uninitialized before training. This allowed the deep learning models to self initialize the weights randomly. Using unsupervised pretraining first, weights can be appropriately initialised to ensure fast arrival to the optimisation point during training.

- **Real-time application**:
  For the models and feature sets considered, there is an average latency of around 0.25 s during prediction phase from the time of feature extraction to the enhanced speech's reconstruction. This is when the testing dataset is readily available. For a real-time application which would involve binarization of the audio recording for storage first to the signal reconstruction post speech enhancement, there is enough potential for optimising the models for reducing latency further.

- **Transfer Learning**:
  Comparison of the performance evaluation for the models considered in the project with the well known neural networks like GoogLeNet , Alexnet etc can also be studied.

Appendix A

Performance Metrics

A.1 STOI

***Short-Time Objective Intelligibility (STOI)**, measures the correlation between the short-time temporal envelopes of a reference (clean) audio signal and a degraded audio signal for speech intelligibility of human speech [20]. The value range of STOI is typically between 0 and 1, 0 being the worst and 1 being the best intelligibility. STOI values can also be considered to be percentage correct.*

A.2 PESQ

***Perceptual Evaluation of Speech Quality (PESQ)** is the standard metric recommended by the International Telecommunication Union (ITU) for analysing quality of a degraded signal with respect to a clean reference signal [21]. PESQ applies an auditory transform to produce a loudness spectrum, and compares the loudness spectra of a clean reference signal and a degraded signal to produce a score in a range of negative 0.5 to 4.5. This score is regarded to be a Mean Opinion Score (MOS). MOS can further be transformed in terms of listening objectivity metrics on a scale from 0 to 5 known as Listening Quality Objectivity (LQO). LQO scale is important as it is a mapping from MOS as per the human auditory response.*

Bibliography

[1] W. contributors, *Cocktail party effect wikipedia the free encyclopedia*, Online; accessed 11-July-2019, 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Cocktail_party_effect.

[2] W. H. Organization. (Mar. 2019). Deafness and hearing loss. fact sheet, [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/deafness-and-hearing-loss.

[3] D. I. N. et al., "The global burden of occupational noise-induced hearing loss", *American journal of industrial medicine*, vol. 48.6, pp. 446–458, 2005.

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE/ACM Trans. Audio, Speech, Sig. Process.*, vol. 27.2, no. 0096-3518, pp. 113–120, 1979.

[5] S. Tamura and A. Waibel., "Noise reduction using connectionist models", *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pp. 553–556, 1988.

[6] C. C. e. a. Awni Hannun. (2014). Deep speech: Scaling up end-to-end speech recognition. in: Cornell university abs/1412.5567, [Online]. Available: https://arxiv.org/abs/1412.5567.

[7] J. C. et al., "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises", *The Journal of the Acoustical Society of America*, vol. 139.5, pp. 2604–2612. 2016.

[8] P.-S. H. et al., "Joint optimization of masks and deep recurrent neural networks for monaural source separation", *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23.12, pp. 2136–2147, 2015.

[9] W. contributors, *Lombard effect, wikipedia the free encyclopedia*, Online; accessed 11-July-2019, 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Lombard_effect.

[10] ——, *Supervised learning, wikipedia the free encyclopedia*, Online; accessed 11-July-2019, 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Supervised_learning.

[11] e. a. Garofolo John S., *Timit acoustic-phonetic continuous speech corpus ldc93s1. web download*, 1993.

[12] D. Snyder, G. Chen, and D. Povey, *Musan: A music, speech, and noise corpus*, arXiv:1510.08484v1, 2015. eprint: 1510.08484.

[13] W. contributors, *Feature (machine learning), wikipedia the free encyclopedia*, Online; accessed 11-July-2019, 2018. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Feature_(machine_learning).

[14] M. Contributors, *Gtcc*, Online; accessed 11-July-2019. [Online]. Available: https://www.mathworks.com/help/audio/ref/gtcc.html.

[15] ——, *Mfcc*, Online; accessed 11-July-2019. [Online]. Available: https://www.mathworks.com/help/audio/ref/mfcc.html.

[16] T. S. C. Hummersone and T. Brooks, *On the ideal ratio mask as the goal of computational auditory scene analysis*, Blind Source Separation, G. R. Naik and W. Wang, Eds. Berlin, Germany, 2014.

[17] W. contributors, *Convolutional neural network, wikipedia the free encyclopedia*, Online; accessed 11-July-2019, 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Convolutional_neural_network.

[18] ——, *Batch normalization, wikipedia the free encyclopedia*, Online; accessed 11-July-2019, 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Batch_normalization.

[19] ——, *Rectifier (neural networks) — wikipedia the free encyclopedia*, Online; accessed 11-July-2019, 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Rectifier_(neural_networks).

[20] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers", *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[21] *Perceptual evaluation of speech quality (pesq), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, ITU-T Recommendation P. 862, 2000.