

Thesis/Project No: CSER-M-19-05

# **Speech Enhancement Using Convolutional Denoising Auto-Encoder**

by

**Shaikh Akib Shahriyar**

A thesis submitted in partial fulfillment of the requirements for the degree of  
Master of Science in Computer Science & Engineering

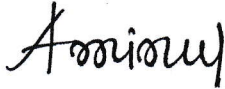


Department of Computer Science and Engineering  
Khulna University of Engineering & Technology  
Khulna 9203, Bangladesh

May, 2019

## Declaration

This is to certify that the thesis work entitled “Speech Enhancement Using Convolutional Denoising Auto-Encoder” has been carried out by Shaikh Akib Shahriyar in the Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna 9203, Bangladesh. The above thesis work or any part of this work has not been submitted anywhere for the award of any degree or diploma.



---

Signature of Supervisor

**Dr. Muhammad Aminul Haque Akhand**

Professor

Dept. of Computer Science and Engineering,  
Khulna University of Engineering & Technology



---

Signature of Candidate

**Shaikh Akib Shahriyar**

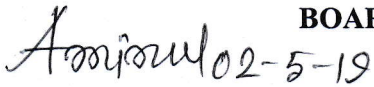

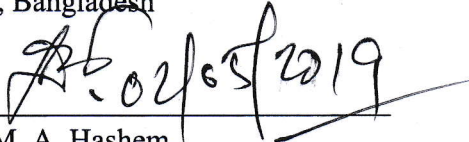
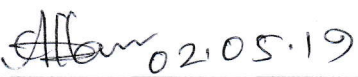
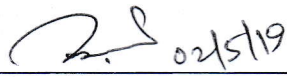
Roll: 1707556

Dept. of Computer Science and Engineering,  
Khulna University of Engineering & Technology

## Approval

This is to certify that the thesis work submitted by Shaikh Akib Shahriyar entitled “Speech Enhancement Using Convolutional Denoising Auto-Encoder” has been approved by the board of examiners for the partial fulfillment of the requirements for the degree of Master of Science in Computer Science & Engineering in the Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna, Bangladesh in May, 2019.

### BOARD OF EXAMINERS

1.  02-5-19  
Dr. Muhammad Aminul Haque Akhand  
Professor  
Dept. of Computer Science and Engineering  
Khulna University of Engineering & Technology  
Khulna, Bangladesh  
Chairman  
(Supervisor)
2.  02-5-19  
Dr. Muhammad Aminul Haque Akhand  
Head  
Dept. of Computer Science and Engineering  
Khulna University of Engineering & Technology  
Khulna, Bangladesh  
Member
3.  02/05/2019  
Dr. M. M. A. Hashem  
Professor  
Dept. of Computer Science and Engineering  
Khulna University of Engineering & Technology  
Khulna, Bangladesh  
Member
4.  02.05.19  
Dr. K. M. Azharul Hasan  
Professor  
Dept. of Computer Science and Engineering  
Khulna University of Engineering & Technology  
Khulna, Bangladesh  
Member
5.  02/5/19  
Dr. Kamrul Hasan Talukder  
Professor  
Computer Science & Engineering Discipline  
Khulna University, Khulna  
Member  
(External)

## **Acknowledgement**

First of all, I would like to thank Almighty for showering all his blessings on me whenever I needed. It is my immense pleasure to express my indebtedness and deep sense of gratitude to my supervisor Dr. Muhammad Aminul Haque Akhand, Professor & Head, Department of Computer Science and Engineering (CSE), Khulna University of Engineering & Technology (KUET) for his continuous encouragement, constant guidance and keen supervision throughout of this study. I learned from Dr. Aminul that persistent effort and faith are undoubtedly the most important assets. I am especially grateful to him for giving me his valuable time whenever I need and always providing continuous support in my effort.

I would also like to express my gratitude to my parents, family members and friends for their patience, support and encouragement during this period.

Last but not least, I am very much grateful to my wife, Khadiza Chowdhury, for her constant understanding, support and love. Without her sacrifice, this dissertation would never have come into existence.

May, 2019

Shaikh Akib Shahriyar

## **Abstract**

Speech signals are complex in nature with respect to other forms of communication media such as text or image. Different forms of noises (e.g., additive noise, channel noise, babble noise) interfere with the speech signals and drastically hamper the quality of the speech. Enhancement of speech signals is a daunting task considering multiple forms of noises while denoising a speech signal. Certain analog noise eliminator models have been studied over the years for this purpose. Researchers have also delved into some machine learning techniques (e.g., artificial neural network) to enhance speech signals. In this study, a speech enhancement system is investigated using Convolutional Denoising Autoencoder (CDAE). Convolutional neural network (CNN) is a special kind of deep neural networks which is suitable for 2D structured input (e.g., image) and CDAE is a CNN based special kind of Denoising Autoencoder. CDAE takes advantages from the 2D structured inputs of the features extracted from speech signals and also considers the local temporal relationship among the features. In the proposed system, CDAE is trained considering features from noisy speech signal as input and clean speech features as desired output. The proposed CDAE based method has been tested on a benchmark dataset, called Speech Command Dataset, and attained 80% similarity between denoised speech and actual clean speech. The proposed system achieved perceptual evaluation of speech quality (PESQ) value of 2.43 which outperformed other related existing methods.

# Contents

	<b>PAGE</b>
Title Page	i
Declaration	ii
Approval	iii
Acknowledgement	iv
Abstract	v
Contents	vi
List of Table	viii
List of Figures	ix
Nomenclature	x
<b>CHAPTER I</b>	
<b>Introduction</b>	1
1.1 Background	1
1.2 Problem Statement	1
1.3 Thesis Objectives	3
1.4 Thesis Contribution	3
1.5 Thesis Organization	3
<b>CHAPTER II</b>	
<b>Literature Review</b>	5
2.1 Noises in Speech Signal	5
2.1.1 Acoustic Noise	5
2.1.2 Thermal Noise and Shot Noise	6
2.1.3 Electromagnetic Noise	6
2.1.4 Electrostatic Noise	6
2.1.5 Channel Distortions	6
2.1.6 Processing Noise	6
2.2 Analog Speech Enhancement	6
2.2.1 Spectral Subtraction	7
2.2.2 Wiener Filtering	8
2.3 Machine Learning based Speech Enhancement	9
2.3.1 Convolutional Neural Network	10
2.3.2 Autoencoder	12
2.3.3 Denoising Autoencoder	13
2.4 Feature of Speech Signal - MFCC	13
2.5 Observation on Existing Methods	16
2.6 Scope of Research	16
<b>CHAPTER III</b>	
<b>Speech Enhancement Using Convolutional Denoising Auto-Encoder</b>	18
3.1 Convolutional Denoising Autoencoder (CDAE)	18
3.2 System Architecture	19
3.2.1 Training Stage	20

	3.2.2	Enhancement Stage	20
	3.3	CDAE Structure for Speech Enhancement	21
	3.4	Noise Consideration	21
	3.5	Evaluation Metrics	21
	3.5.1	Structural Similarity Index Measure	22
	3.5.2	Perceptual Evaluation of Speech Quality	22
<b>CHAPTER IV</b>		<b>Experimental Studies</b>	25
	4.1	Experimental Setup	25
	4.2	Input Dataset Preparation	25
	4.3	Noise List	25
	4.4	MFCC Feature Extraction	26
	4.5	Experimental Results and Analysis	27
<b>CHAPTER V</b>		<b>Conclusions</b>	29
	6.1	Contribution	29
	6.2	Future Work	29
<b>PUBLICATION FROM THE THESIS</b>			30
<b>REFERENCES</b>			31

## LIST OF TABLES

<b>Table No</b>	<b>Caption of the Table</b>	<b>Page</b>
4.1	List of Noises used for training	24



## LIST OF FIGURES

<b>Figure No</b>	<b>Caption of the Figure</b>	<b>Page</b>
2.1	Enhancement of Speech Signal by Spectral Subtraction.	7
2.2	Enhancement of Speech Signal by Wiener Filtering	9
2.3	Basic Structure of a Convolutional Neural Network.	10
2.4	Basic Structure of an Autoencoder.	12
2.5	Basic Structure of a Denoising Autoencoder.	13
2.6	Filter Space of MFCC	14
2.7	MFCC Coefficients of a Speech Signals.	15
3.1	Basic Structure of a CDAE.	18
3.2	Overall Architecture of Proposed Speech Enhancement System.	19
3.3	Structure of the proposed CDAE for Speech Enhancement System.	21
3.4	Overview of the basic philosophy used in PESQ.	23
4.1	Training and Test Sets Similarity for Different Iterations with Batch Size 100.	25
4.2	Comparison of PESQ Values of Different Speech Enhancement Models.	26

## Nomenclature

AE	Autoencoder
ASR	Automatic Speech Recognition
CDAE	Convolutional Denoising Autoencoder
DAE	Denoising Autoencoder
DNN	Deep Neural Network
FCN	Fully Convolutional Network
FNN	Feed forward Neural Network
MFCC	Mel Frequency Cepstral Coefficients
MMSE	Minimum Mean Square Error
NMF	Nonnegative Matrix Factorization
PESQ	Perceptual Evaluation of Speech Quality
RNN	Recurrent Neural Network
SSIM	Structural Similarity Index Measure
SVM	Support Vector Machines

# CHAPTER I

## Introduction

Enhancement of speech signals is one of the most complex tasks faced by researchers in speech processing field. This chapter provides an overview of problems generated by noises in speech signals, describes the problem statement, point out the objectives of the thesis, provides a short brief on thesis contribution along with the thesis organization.

### 1.1 Background

Speech signals are complex in nature with respect to other forms of communication media such as text or image. Speech signals are continuously corrupted with different sort of noises in various speaking conditions. Different forms of noises (e.g., additive noise, channel noise, babble noise) interfere with the speech signals and drastically hamper the quality of the speech in the speech signals. To enhance any noisy speech signal and turn it into a clean speech, the noise has to be tackled with immense expertise. Thus, enhancement of speech signals is a daunting task considering multiple forms of noises while denoising a speech signal.

### 1.2 Problem Statement

Enhancement of speech from noisy speech signals have become a prominent research area over the years. A speech enhancement system is required to denoise the noisy speech signals and convert them into clean speech signals. Earlier, some noise reduction models based on analog procedures have been developed by incessant research on speech enhancement field which provided some feasible solutions. Lim and Oppenheim compressed the width of the bandwidth to enhance noisy speech which slightly cleaned the noisy speech signal [1]. Using Minimum Mean Square Error (MMSE), Ephraim and Malah proposed a speech enhancement method which estimates small burst of spectral amplitudes [2]. Most of these methods work upon either signal subspace approach or Signal-to-Noise estimation method [3], [4].

On the other hand, Machine Learning techniques usually take a different route than the signal subtraction methods. ML models try to learn a specific mapping of speech signal features between noisy and clean speech. The feature learning procedure of machine learning models differs in many ways. Complex form of Feed Forward Neural Network (FNN) have been used for denoising noisy speech signals by Osako et al. [5]. Wang et al. trained linear Support Vector Machines (SVMs) to properly separate noise features from speech signals [6]. To enhance speech signals embedded in nonstationary noise, Sameti et al. used Hidden Markov Models which used mixture components in a flexible manner [7]. Fu et al. proposed a fully convolutional network (FCN) for raw waveform based speech enhancement [8]. The FCN used raw speech signals as input and produced cleaner speech signals as output. Wilson et al. proposed Nonnegative Matrix Factorization (NMF) for speech denoising considering some prior assumptions [9].

Deep Neural Network (DNN) based speech enhancement has also become popular in recent times [10], [11] and it gives much better result than traditional analog models. Han et al. learned the mapping of spectral region of speech through neural network [12]. Weighted denoising autoencoders have been used to clean noisy speech in [13]. Recurrent Neural Networks (RNN) have also been considered for denoising noisy speech by Osako et al. [14]. In traditional models, like Negative Matrix Factorization (NMF) [9], computation process is much heavier than DNN. Thus, DNN based approach is a much effective choice for speech enhancement. However, DNN based approach could not characterize the temporal structures of any speech and does not take into consideration any locally available connection between features of that speech signal.

The issue of considering local temporal features of a speech signal can be partially solved by using Convolutional Neural Network (CNN) [15]. CNN can be also coupled with DNN and HMM for more diverse speech enhancement [16]. As CNN takes 2D inputs, it learns the temporal features through its convolution layers and creates a multi-level hierarchical abstraction through sub-sampling layers. CNN based model takes input of the spectral images of any speech signal and enhances speech through spectral noise removal process [5], [15]. The model proposed by Tsao and Park et al. does not take into consideration of any other important features of a speech signal other than spectrogram of signals. In this thesis, Mel Frequency Cepstral Coefficients (MFCC), one of the most prominent technique for feature representation of a speech signal [17] is being considered.

### **1.3 Thesis Objectives**

The key objective of this research is to investigate Convolutional Denoising Autoencoder (CDAE) and its ability to enhance the speech signals. To reach the goal the study has been carried out with the following specific objectives:

- MFCC Feature extraction from speech signals.
- Design CDAE for speech enhancement.
- Performance comparison with other prominent works (FNN, CNN, RNN) to identify effectiveness of CDAE for enhancing speech signal.

### **1.4 Thesis Contribution**

This study investigates a speech enhancement system which takes into consideration of the local temporal relationship between MFCC features and enhances speech accordingly by using Convolutional Denoising Autoencoder (CDAE). Proposed CDAE based approach takes the MFCC features as input which is formatted in 2D structure and maintains the local temporal relationship between important features of a speech signal. The system cleans different types of acoustic noises from the noisy input speech signals and produces clean speech signal as output. The proposed speech enhancement system overcomes the bottlenecks present in 1D structured feature input and enhances the quality and increase the intelligibility of speech signals.

### **1.5 Thesis Organization**

The thesis is organized in five chapters.

Chapter I provides introductory discussion on background and motivation of speech enhancement, thesis objectives and thesis organization.

Chapter II provides overview of speech signals, brief discussion on existing methods to enhance speech signals, observation on some preliminary deep learning methods and scope of the research.

Chapter III describes the proposed method of speech enhancement using CDAE model in details.

Chapter IV contains experimental studies. In this chapter experimental setup, environment, input data preparation, experimental results and analysis are discussed.

Chapter V provides concluding remarks and possible future research directions.

## **CHAPTER II**

### **Literature Review**

Speech enhancement is an exciting field where a lot of techniques are already available and in widespread use, but plenty of challenges still remain. This chapter provides overview of some earlier speech enhancement approaches such as Spectral Subtraction Method, Wiener Filtering and also some Machine Learning and Deep Learning methods such as CNN, Autoencoder (AE), Denoising Autoencoder (DAE) along with the MFCC feature extraction process. It also includes brief observation on existing methods for speech enhancement and scope of research.

#### **2.1 Noises in Speech Signal**

In signals, noise is counted as inadmissible. Usually, signals which uses undirected wireless transmission are the most affected by noises. These undesired frequency signals affect quality of those transmitted signals. Thus, noise should be handled with care as it can cause severe amount of error in any system, such as speech recording system, voice guided mobile applications etc.

In practical life, we experience a lot of background noises and other forms of distortion while hearing any speech. To properly hear the actual speech signal, speech enhancement is needed for attenuating the noise. The removal of noise from speech signals can be modeled properly by a speech enhancement system by investigating the types of noises and teaching the system to learn these noises from the input speech signals. Noise can be categorized [18] into the following different categories:

##### **2.1.1 Acoustic Noise**

This type of noise is produced from different set of sounds which are generated from motion of objects, vibrations in specific directions and collision. Acoustic noise is also generated by generators like car motion, electronic equipment such as air conditioners and computers. Also, it is generated by different undirected and unguided surrounding activities in nature. We have considered various acoustic noises in this thesis.

### **2.1.2 Thermal Noise and Shot Noise**

Thermal noise is generated by heat only. All the equipment and semiconductor devices operated at the room temperature. Thermal noise results from uneven motion of energized particles in the electrical conductor because of temperature. Thermal noise is there in all conducting medium and is produced without any application of electromotive force. In contrast because of heavy undirected motion of electrical current in the conductor shot noise is there.

### **2.1.3 Electromagnetic Noise**

This noise can be generated at all frequencies in the band and wherever long-distance frequency travel and communication takes place, this noise is present. All the electronics equipment working on the radio frequency.

### **2.1.4 Electrostatic Noise**

The most important sources of electrostatic noise are fluorescent lighting. It is not important whether flow of current is there or not

### **2.1.5 Channel Distortions**

Whenever there is relative motion between transmitter and receiver it causes fluctuation in the received signal and it causes fading of the signal. Because of generated fading signal strength becomes weak and because of that quality of produced signal will be degraded.

### **2.1.6 Processing Noise**

Noise which results from internal D/A processing of signal like phenomena of quantization in which quantization errors are generated. Moreover also because of erroneous channels data packets are lost in the system.

## **2.2 Analog Speech Enhancement**

Analog Speech enhancement methods can be classified into different type of categories such as Spectral Subtraction method, Wiener Filtering method and its different variations. The performance of speech enhancement methods differs in case of speech quality, intelligibility, noise suppression etc. The choice of speech enhancement methods also depends on the constraints and assumptions of each method about noise and speech environment. The computational complexity of each method also plays a vital role for their selection as a



speech enhancement-oriented application. Among various speech enhancement methods, two prominent methods named “Spectral Subtraction method” and “Wiener Filtering method” will be discussed in the next Section.

### 2.2.1 Spectral Subtraction

Estimation of noise and removal of such noise from speech signal is the basic principle followed in Spectral Subtraction Method. In this method, speech and background noise is considered as uncorrelated. The uncorrelated noise source  $N(z)$  can be considered as additive noise onto a speech signal  $S(z)$  which formulates an observation,

$$X(z) = S(z) + N(z)$$

The SSM estimates  $S(z)$  where  $X(z)$  is unknown.  $|\hat{N}(z)|^2$  is a noisy energy estimate, but  $N(z)$  is not known. Thus, a subtraction is performed as

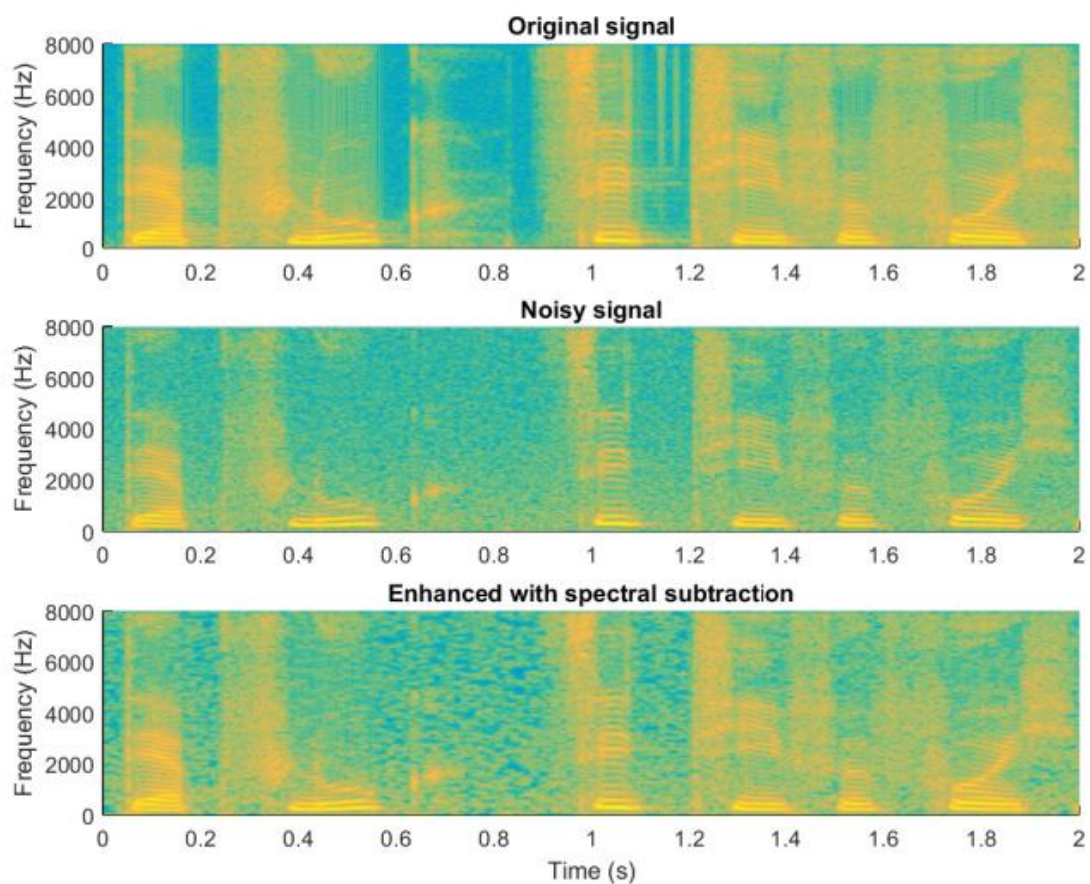


Figure 2.1: Enhancement of Speech Signal by Spectral Subtraction [19].

$$|\hat{S}(z)|^2 = |\hat{X}(z)|^2 - |\hat{N}(z)|^2 \text{ where}$$

$$|S(z)| = \sqrt{|X(z)|^2 - |\hat{N}(z)|^2}$$

The phase of the observation  $\angle X(z)$  is kept as the speech signal's phase such that

$$\angle \hat{S}(z) = \angle X(z)$$

The estimated speech signal will be

$$|\hat{S}(z)| = \angle X(z) \sqrt{|X(z)|^2 - |\hat{N}(z)|^2}$$

If the speech energy estimate become negative the following limiting function is followed

$$|\hat{S}(z)|^2 = \begin{cases} |X(z)|^2 - |\hat{N}(z)|^2, & |X(z)|^2 > |\hat{N}(z)|^2 \\ 0, & \text{Otherwise} \end{cases}$$

If the method overestimated  $|\hat{N}(z)|^2$ , then a bit less is subtracted. If the method underestimated  $|\hat{N}(z)|^2$ , then further subtraction is carried out. The spectral subtraction method gives a biased estimate which often removes too much speech energy. Figure 2.1 gives a clear idea of speech enhanced by spectral subtraction method on a noisy input signal. Boll et al. proposed a spectral subtraction based approach to remove acoustic noise [20]. The method proposed by Boll et al. suppresses stationary noise from speech signals which required huge computations to subtract spectral noise. Berouti et al. proposed a similar approach for enhancement of speech using spectral floors [21].

### 2.2.2 Wiener Filtering

Wiener Filters works similarly to Spectral Subtraction method in many ways. It minimizes the squared error between the original speech signal  $S(z)$  and an estimate of that signal  $\hat{S}(z)$ .

$$\min_{\hat{S}(z)} E \left[ |S(z) - \hat{S}(z)|^2 \right]$$

Thus, Wiener filter estimates clean speech by the following equation

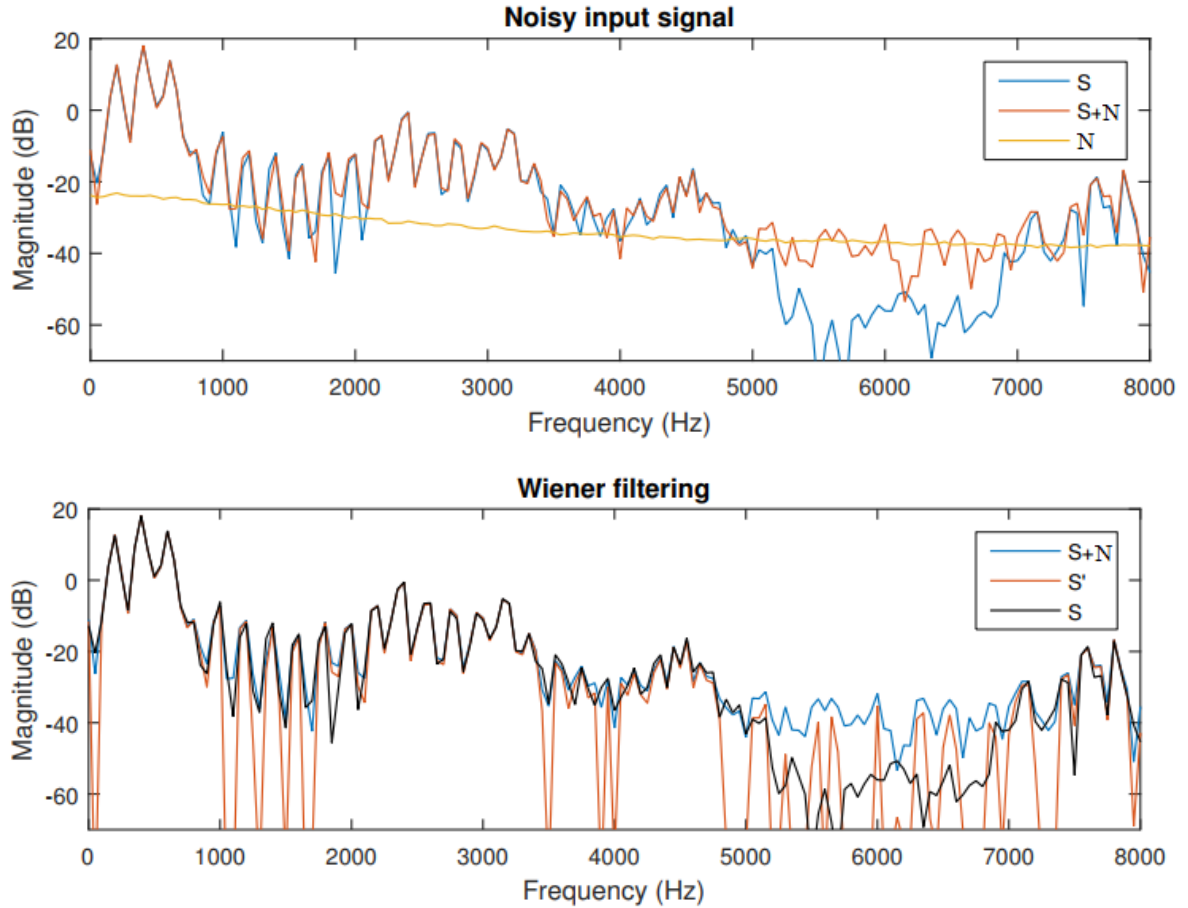


Figure 2.2: Enhancement of Speech Signal by Wiener Filtering [24].

$$\hat{S}(z) = X(z) \left[ \frac{|X(z)|^2 - |\hat{N}(z)|^2}{|X(z)|^2} \right]$$

Speech enhancement by Wiener filter can be seen in Fig. 2.2. Leonid et al. proposed a nonlinear variation based noise removal algorithm where wiener filters were explicitly applied [22]. Xu et al. proposed domain filters to transform wavelet and application of wiener filtering to suppress selective noise in signals [23].

### 2.3 Machine Learning based Speech Enhancement

Machine learning based approached for speech enhancement requires feature extraction from speech signals along with the mapping of these features from noisy speech to clean speech. The feature considered in this thesis for speech enhancement is MFCC. ML Methods such as Convolutional Neural Network [15], Autoencoders [25], Denoising Autoencoders

[26] have been used over the years for speech enhancement. The methods along with the MFCC feature extraction process is discussed in the next Section.

### 2.3.1 Convolutional Neural Network

CNN was first introduced by Lecun et al. for document recognition along with time series data such as speech signal [27], [28]. The structure of CNN can be seen in Fig. 2.3. The CNN architecture was fundamental, in particular the insight that image features are distributed across the entire image, and convolutions with learnable parameters are an effective way to extract similar features at multiple location with few parameters. Thus, the idea of collecting features from 2D structures generated and it is being implemented till date.

The network consists of some different operations. They are

1. **Convolution:** It is a function derived from two given functions by integration which expresses how the shape of one is modified by the other. Convolution is the first layer to extract features from an input image.

$$(f * g) \stackrel{def}{=} \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau$$

Convolution preserves the relationship between pixels by learning image features using small squares of input data. It is a mathematical operation that takes two inputs such as image matrix and a filter or kernel.

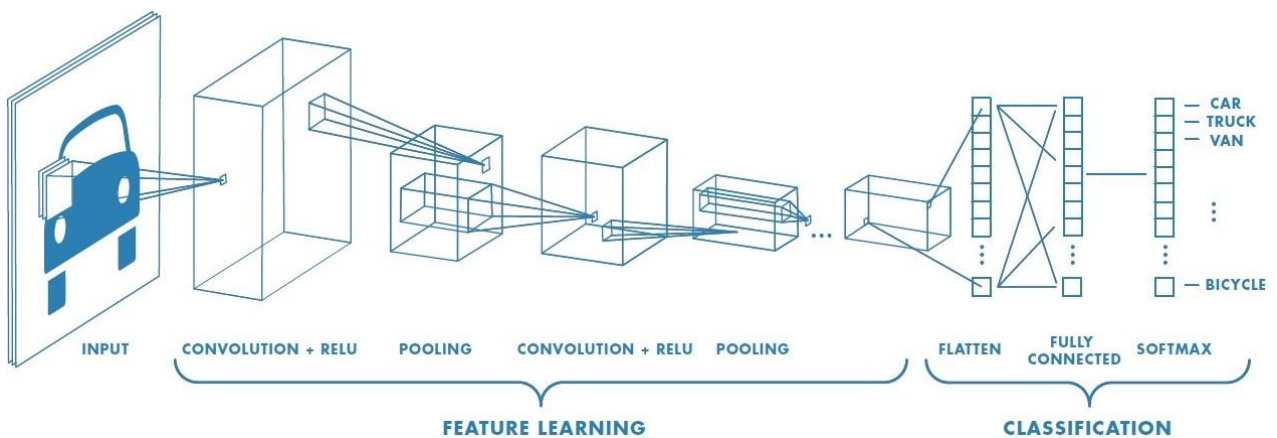


Figure 2.3: Basic Structure of a Convolutional Neural Network [29].

2. **Pooling Layer:** Pooling layers section would reduce the number of parameters when the images are too large. Spatial pooling also called subsampling or downsampling which reduces the dimensionality of each map but retains the important information. Spatial pooling can be of different types:
  - Max Pooling
  - Average Pooling
  - Sum Pooling
  
3. **Non-Linearity (ReLU):** ReLU stands for Rectified Linear Unit for a non-linear operation. ReLU's purpose is to introduce non-linearity in CNN. Since, the real-world data need CNN to learn non-negative linear values. There are other non-linear functions such as tanh or sigmoid can also be used instead of ReLU.
  
4. **Fully Connected Layer:** The matrix output of the previous layers is flattened and fed into the fully connected layer as similar to neural networks.
  
5. **Softmax:** It is a generalised version of the logistic function that squashes arbitrary values of the variables to a real value between 0 to 1. The equation for calculating softmax is

$$f_i(z) = \frac{e^{z_j}}{\sum_k e^{z_k}}$$

It helps to classify the input image into its corresponding class by providing the maximum real value to a corresponding class in between 0 to 1.

As CNN is an ideal model for processing images and retrieving of temporal features from images, it was used on spectrogram of speech signals by Fu et al. [30]. It used segmental signal-to-noise ratio for checking the effectiveness of CNN on spectrogram. Kounovsky et al. proposed a similar structure for single channel speech enhancement using CNN [31]. Pandey et al. proposed a temporal convolutional module additional to general CNN which enhanced speech better than ordinary CNN [32].

### 2.3.2 Autoencoder

An autoencoder simply changes the representation of data from one layer to another [25]. The process is simply denoted as ‘latent representation’. Suppose one needs to map  $\alpha$  to  $\gamma$ . It takes an input  $\alpha$  and then maps the input data to the first layer of latent representation  $\beta$  by following a deterministic function. The input can be the  $\alpha \in [0,1]^a$  which can be mapped to a hidden layer as  $\beta \in [0,1]^a$  as

$$\beta = f(W\alpha + a)$$

where  $f$  is a nonlinear function. The latent representation  $\beta$  is then mapped to the final representation  $\gamma$  which has the same dimension as  $\alpha$  using Eq. (2).

$$\gamma = f(W'\alpha + a')$$

It uses backpropagation to learn the corresponding weights  $W$ ,  $W'$  using different loss functions. Fig. 2.4. depicts the basic structure of an autoencoder. Autoencoders are vastly used for unseen noise estimation [26] by modelling them as denoising autoencoders but they lack the temporal feature consideration.

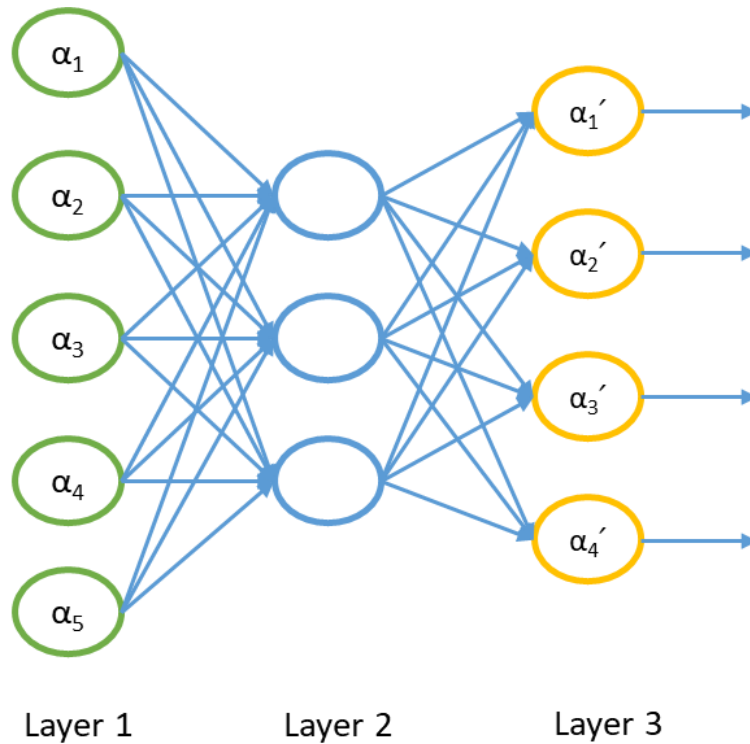


Figure 2.4: Basic Structure of an Autoencoder.

### 2.3.3 Denoising Autoencoder

Denoising autoencoder [33] is a special type of autoencoder which makes the model learn the procedure of reconstruction of input provided its noise mixed version. It forces the denoising autoencoder to predict missing values to reconstruct a clean version of the noisy input data. It takes any length of 1D input and incorporates noise within it which forces the system to learn from the noisy data using the activation functions  $f$ . The mapping of these noisy data to clean data is completed by the latent representation of features as depicted in Fig. 2.5. Here,  $\alpha$  is the raw input,  $\alpha'$  is the corrupted input,  $\beta$  is the latent representation and  $\gamma$  is the reconstruction of the raw input.

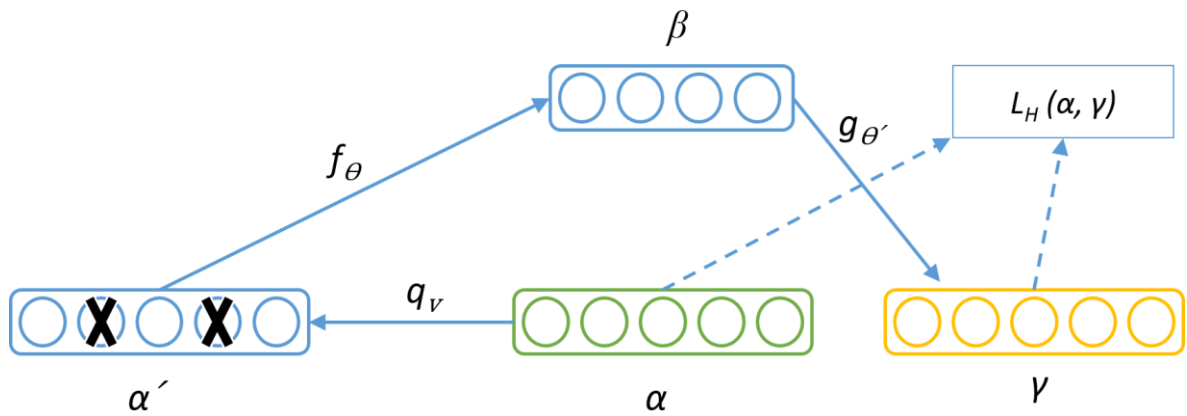


Figure 2.5: Basic Structure of a Denoising Autoencoder.

DAE is unable to consider 2D inputs and require to transform 2D to 1D and might not be effective for all the cases where 2D inputs plays a vital role. DAE performs relatively well than Restricted Boltzmann machines (RBM) [34]. Some multi-channel features are also considered for DAE which was proposed by Araki et al [35]. Xu et al. proposed a Ensemble modelling of DAE which restored spectrums in speech signals [36].

### 2.4 Feature of Speech Signal - MFCC

MFCC was introduced in 1980's by Davis and Mermelstein and crowned as the state-of-the-art in extracted features from audio signals [17]. It mimics the vocal envelopes in the form of specific filters, which can extract the uttered phonemes accurately from the speech.

Mel-Frequency analysis of speech is based on human perception experiments. It is observed that human ear acts as a filter. It concentrates on only certain frequency components. Our ear has cochlea which basically has more filters at low frequency and very few filters at higher frequency. This can be mimicked using Mel filters. These filters are non-uniformly spaced on the frequency axis. More filters in the low frequency regions, a smaller number of filters in high frequency regions as can be seen in Fig 2.6.

MFCC can be calculated through the following steps:

1. **Frame Blocking:** The input speech signal is segmented into frames of 20~30 ms with optional overlap of 1/3~1/2 of the frame size. Usually the frame size (in terms of sample points) is equal to power of two in order to facilitate the use of FFT. If this is not the case, we need to do zero padding to the nearest length of power of two. If the sample rate is 16 kHz and the frame size is 320 sample points, then the frame duration is  $320/16000 = 0.02$  sec = 20 ms. Additional, if the overlap is 160 points, then the frame rate is  $16000/(320-160) = 100$  frames per second.

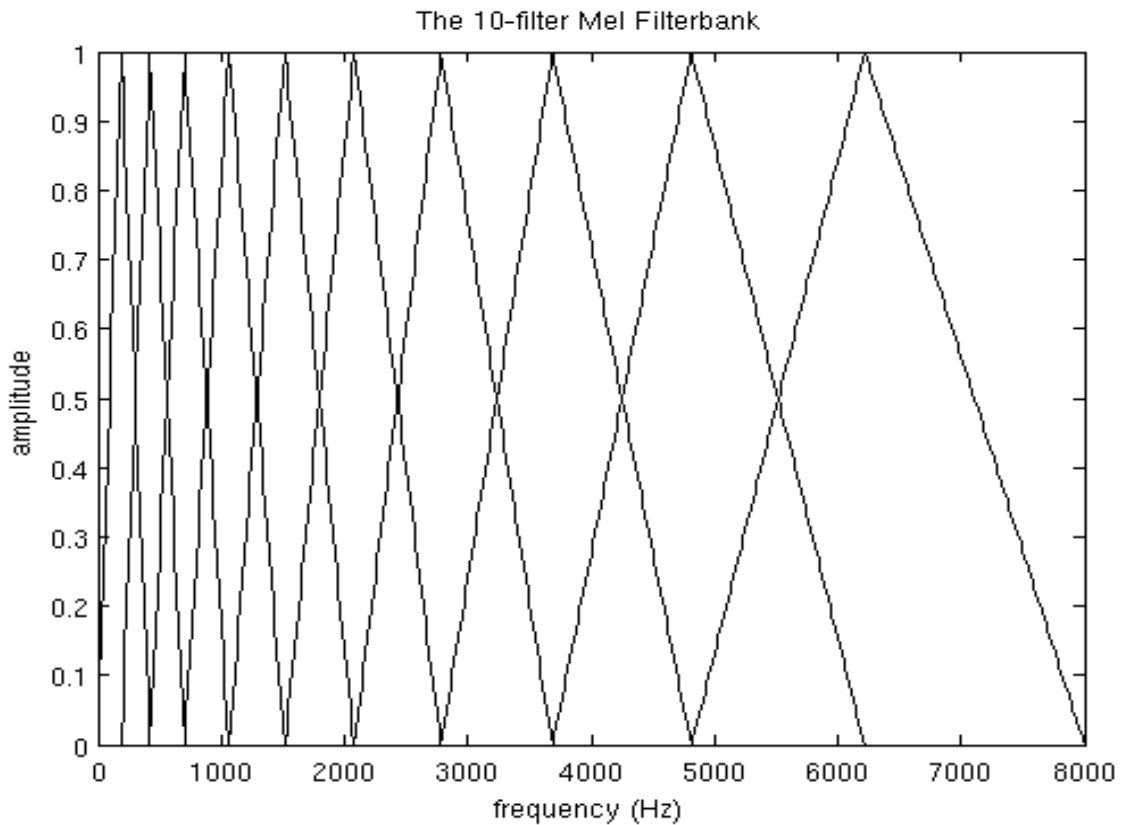


Figure 2.6: Filter Space of MFCC [37].



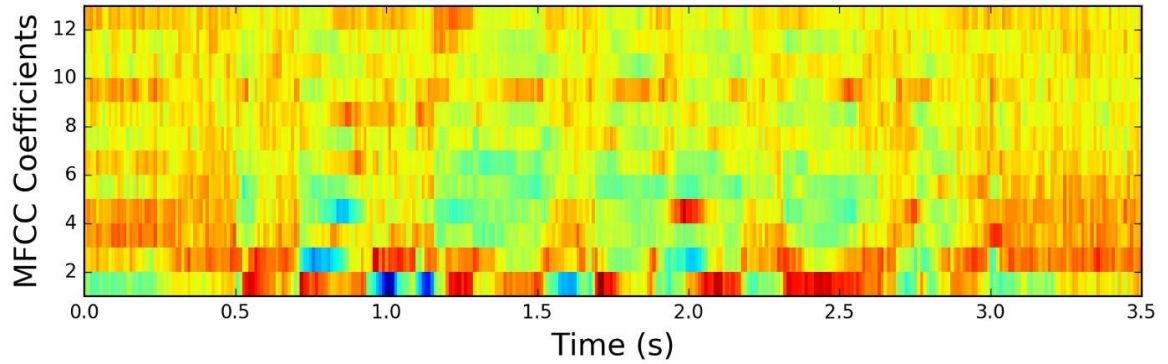


Figure 2.7: MFCC Coefficients of a Speech Signals [38].

2. **Hamming Windowing:** Each frame has to be multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame (to be detailed in the next step). If the signal in a frame is denoted by  $s(n)$ ,  $n = 0, \dots, N-1$ , then the signal after Hamming windowing is  $s(n) * w(n)$ , where  $w(n)$  is the Hamming window defined by:

$$w(n, \alpha) = (1 - \alpha) - \alpha \cos\left(\frac{2\pi n}{N - 1}\right), 0 \leq n \leq N - 1$$

Different values of  $\alpha$  corresponds to different curves for the Hamming windows

3. **Fast Fourier Transform (FFT):** Spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore, we usually perform FFT to obtain the magnitude frequency response of each frame. When we perform FFT on a frame, we assume that the signal within a frame is periodic, and continuous when wrapping around. If this is not the case, we can still perform FFT but the in-continuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response.
4. **Triangular Bandpass Filters:** We multiple the magnitude frequency response by a set of 20 triangular bandpass filters to get the log energy of each triangular bandpass filter. The positions of these filters are equally spaced along the Mel frequency, which is related to the common linear frequency  $f$  by the following equation:

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

5. **Log energy:** The energy within a frame is also an important feature that can be easily obtained. Hence, we usually add the log energy as the 13rd feature to MFCC.
6. **Discrete Cosine Transform (DCT):** In this step, we apply DCT on the 20-log energy  $E_k$  obtained from the triangular bandpass filters to have  $L$  mel-scale cepstral coefficients. The formula for DCT is shown next.

$$C_m = \sum_{k=1}^N \cos\left[m * (k - 0.5) * \frac{\pi}{N}\right] * E_k, m = 1, 2, \dots, L$$

where  $N$  is the number of triangular bandpass filters,  $L$  is the number of mel-scale cepstral coefficients. After applying the steps onto a speech signals, MFCC features can be retrieved as depicted in Fig. 2.7. Here 12 MFCC Coefficients are collected from a speech signal.

## 2.5 Observation on Existing Methods

Analog models such as Wiener filtering and spectral subtraction techniques suffers greatly in case of enduring noise and it has an apparent effect on the cleaned speech signal. With respect to these analog models, Machine Learning methods learn the pattern of the noises and try to perfect the output of the speech enhancement model accordingly. Hidden Markov Models used mixture components in a flexible manner [7]. Some prior assumptions are considered for speech denoising by Nonnegative Matrix Factorization (NMF) [9]. In case of Deep Learning methods, Though AE and DAE doesn't capture the local temporal features, CNN considers local temporal features [15]. CNN based model takes input of the spectral images of any speech signal and enhances speech through spectral noise removal process [5], [15], [31]. Thus, some deep learning models based on Convolutional architecture which has denoising ability such as CDAE, can be investigated for enhancing speech signals. To the best of our knowledge, CDAE has not been used to enhance speech signals.

## 2.6 Scope of Research

From the observation of existing methods, it is clear that there is a scope of enhancing speech signals quality by considering the local temporal features of speech signals. Also, CDAE seems to be promising for enhancing speech signals. As CDAE has not been used to enhance the quality of speech signals, there is a scope of utilizing CDAE for enhancing speech

signals. In this study, MFCC has been considered as the major feature from speech signals and CDAE is being used for enhancing and improving the quality and intelligibility of the speech signals.

## CHAPTER III

### Speech Enhancement Using Convolutional Denoising Auto-Encoder

This chapter presents the proposed Convolutional Denoising Autoencoder approach for Speech Enhancement task. Section 3.1 provides a short brief over CDAE, Section 3.2 provides a details overview of the system architecture, Section 3.3 describes the structure of the proposed CDAE model and Section 3.4 explains about the noises used in this thesis work. Section 3.5 describes about the evaluation metrics of the system.

#### 3.1 Convolutional Denoising Autoencoder (CDAE)

CDAEs consist of the same principle of autoencoders with the added structure of convolutional encoding and deconvolutional decoding layers [39]. CDAE takes 2D structural inputs into consideration and reconstructs the inputs into the same 2D structure using encoder and decoder layers as can be seen in Fig. 2. Thus, CDAE is better suited for speech enhancement because of the learned MFCC features from speech signals are in 2D format. It is similar to DAEs except the fact that weights are shared all over the network which maintains the relationship of local spatiality. As shown in the following equation, for any  $i$ th feature map,

$$h^i = f(W^i * \alpha + a^i)$$

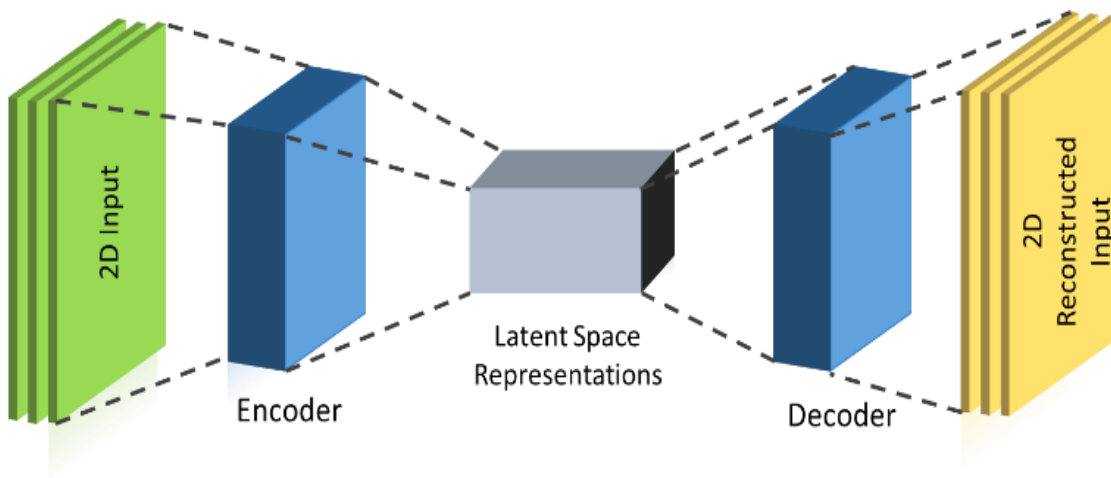


Figure 3.1: Basic Structure of a CDAE.

where \* means convolution process and the bias value is shared over the whole map. By using the bias, the reconstruction can be stated as in the following equation.

$$\beta = s \left( \sum_{i \in J} h^i * W^i + b \right)$$

In the equation,  $b$  is a bias for every input,  $s$  is an activation function,  $J$  represents a set of latent feature maps and  $W$  is the flipped weight. The overall learning of these parameters is conducted by backpropagation.

### 3.2 Proposed System Architecture

The proposed CDAE based speech enhancement system is demonstrated in Fig. 3.2. The system is divided into two stages namely, training stage and enhancement stage.

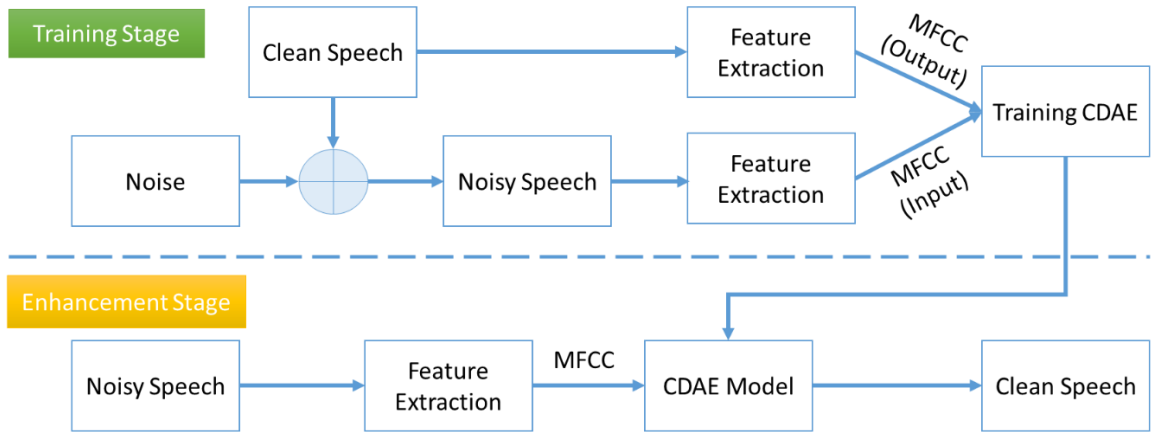


Figure 3.2: Overall Architecture of Proposed Speech Enhancement System.

#### 3.2.1 Training Stage

Firstly, Noise is randomly chosen from a list of acoustic noises which are downsampled to match the sample rate of the input clean speech signals. These noises are additive in nature which creates a distortion in the clean speech signal which can be seen in Fig. 3.2.

Secondly, MFCC features are extracted from the noisy speech signals as well as clean speech signals. These features are stored as a pair which contains Mel-frequency Cepstrum Coefficients of both clean and noisy version of a single utterance. Each tuple contains numeral representation of a clean speech signal and its noisy variant. Combining these tuples create the training dataset which is fed into the proposed Convolutional Denoising

Autoencoder model. The CDAE model is fed the MFCC feature of the noisy variant of a speech signal whereas the MFCC feature of the clean variant of that speech signal is the set as the desired output. The model iterates through multiple epochs to learn the mapping between noisy input speech signal and desired clean speech signal. After the CDAE model trains for a set number of epochs, achieving acceptable train set accuracy, it becomes prepared to be applied on the test set of noisy speech signals.

### **3.2.2 Enhancement Stage**

In the enhancement phase, Noisy speech signals are considered from the test set. These speech signals contain noises which have affected their quality and intelligibility. Firstly, MFCC features are extracted from these noisy speech signals. As the CDAE model is trained with a target to reconstruct noisy speech to clean speech, the extracted features from the noisy speech signals are fed into the trained CDAE model. The model reconstructs the noisy speech signal features into clean speech features. These reconstructed features are called denoised speech features. They are MFCC features which represents the denoised cleaned speech.

Finally, the denoised speech features and the actual clean speech features are measured with evaluation metrics to calculate the effectiveness and performance of the proposed speech enhancement system.

### **3.3 CDAE Structure for Speech Enhancement**

The architecture of the CDAE model can be seen in Fig. 3.3. The input matrix was  $16 \times 16$  in size containing MFCC features from noisy audio signals in the training data. A convolution layer is first introduced with the input size of  $16 \times 16$  and an output size of  $16 \times 16$  along with 64 kernels. Next, a maxpooling layer was introduced with a kernel size of  $2 \times 2$ , which gives the corresponding output of  $8 \times 8$ .

The convolution layer and maxpooling layer is combinedly represented as Encoder 1 in Fig. 3.3. Another convolution layer and maxpooling layer sandwich, which is called Encoder 2 in Fig. 3.3, is put together after the max pooling layer with the same number of kernels thus it gets the output to  $4 \times 4$  size feature matrices. As a result, the encoder portion of the network, converts the  $16 \times 16$  features into a  $4 \times 4$  feature representation.

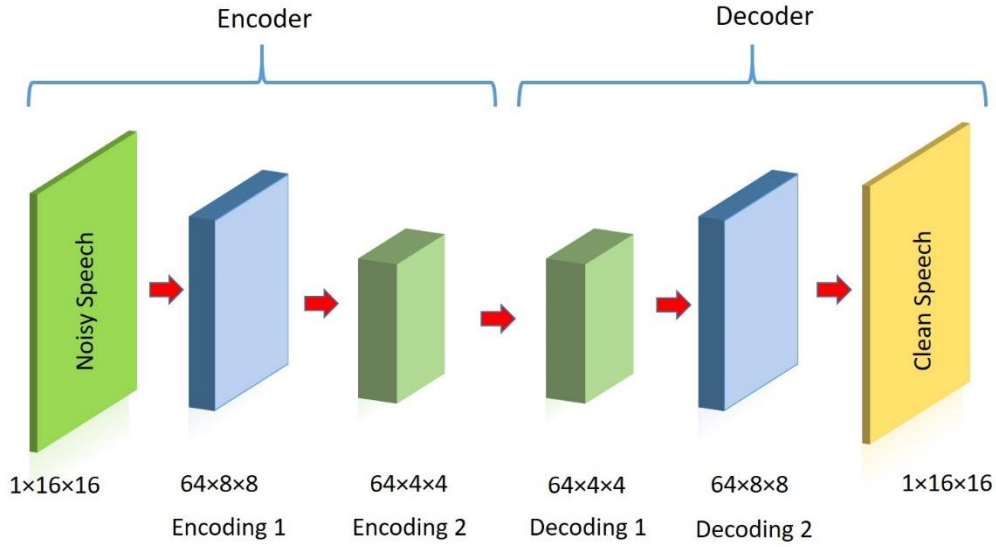


Figure 3.3: Structure of the proposed CDAE for Speech Enhancement System.

The decoder section of the CADE consists of similar layered sandwiches such as Decoder 1 and Decoder 2 of deconvolution layers and upsampling layers, keeping the same ratio to restore the features from 4x4 feature matrices to 16x16 feature matrices. Finally, the fully connected linear layer produced the denoised audio speech features, which can be referred as clean speech.

### 3.4 Noise Consideration

Different types of noise (e.g., car noise, traffic noise, white noise) which were freely available online [40] were added to the utterances to produce noisy speech data from the clean data of the dataset. Each noise has a different set of intensity which can be measured in decibels (dB). Although White noise, Pink Noise and High Frequency Channel Noise doesn't have distinct intensity levels, they were added to the noise list as they are some of the effective acoustic noises. Two or three types of noises were randomly chosen to add to make each utterance noisy. Thus, training data contains enough diversity with different types of noises and mimics real life noisy environment.

### 3.5 Evaluation Metrics

Two types of measures are considered in this study for evaluating the performance of the proposed system.

#### 3.5.1 Structural Similarity Index Measure

Structural Similarity Index Measure (SSIM) [41] was used between the denoised speech signal and clean speech signal to compute the similarity and accuracy of the proposed system. As SSIM is a perception-based model which relies on the structural information to measure similarity. The base idea is that each component in 2D structure has strong dependencies with spatially close components. These dependencies play a vital role in carrying significant information about the 2D structure. Eq. (5) shows the SSIM calculation procedure.

$$SSIM(p,q) = [X(p,q)]^\alpha [Y(p,q)]^\beta [Z(p,q)]^\gamma$$

The SSIM formula is based on three comparison measurements between the samples of  $p$  and  $q$ . In the equation, X, Y and Z are structural components calculated from the mean, standard deviation and covariance from the denoised speech features and the clean speech features. Here,  $\alpha$ ,  $\beta$  and  $\gamma$  are greater than 0 and maintains relative significance of SSIM.

#### 3.5.2 Perceptual Evaluation of Speech Quality

Perceptual Evaluation of Speech Quality (PESQ) [42] is considered for evaluation of the quality of the denoised speech signal. PESQ is vastly used in telecommunication sector as it is one of the global standard quality measure metrics used by International Telecommunication Union (ITU) [43]. The standard used by ITU is stated as “ITU-T P.862.3”. The test is an objective test which is designed to mimic the subjective test on the quality of speech signals. It characterizes the listening quality of any speech signal on the perspective of human perceivability. PESQ compares an original signal  $X(t)$  with a degraded signal  $Y(t)$  that is the result of passing  $X(t)$  through a communications system. The output of PESQ is a prediction of the perceived quality that would be given to  $Y(t)$  by subjects in a subjective listening test.



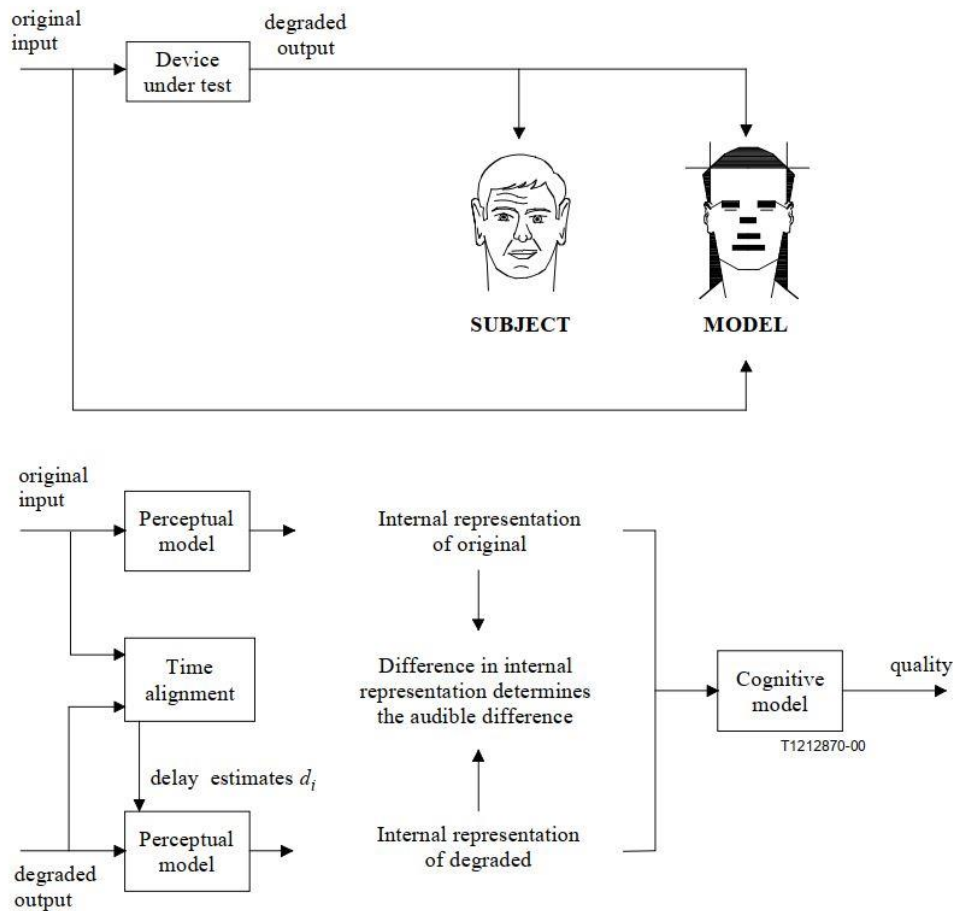


Figure 3.4: Overview of the basic philosophy used in PESQ [43].

In the first step of PESQ a series of delays between original input and degraded output are computed, one for each time interval for which the delay is significantly different from the previous time interval. For each of these intervals a corresponding start and stop point is calculated. The alignment algorithm is based on the principle of comparing the confidence of having two delays in a certain time interval with the confidence of having a single delay for that interval. The algorithm can handle delay changes both during silences and during active speech parts. Based on the set of delays that are found PESQ compares the original (input) signal with the aligned degraded output of the device under test using a perceptual model, as illustrated in Fig 3.4. The key to this process is transformation of both the original and degraded signals to an internal representation that is analogous to the psychophysical representation of audio signals in the human auditory system, taking account of perceptual frequency (Bark) and loudness (Sone). This is achieved in several stages: time alignment,

level alignment to a calibrated listening level, time-frequency mapping, frequency warping, and compressive loudness scaling.

PESQ analyzes the speech signals completely by using a full-reference algorithm which checks the speech signal sample-by-sample. It is also used for quality assessment task of any telecommunication network where quality of speech signals has great importance. The PESQ scores the quality of any speech signal with a range from -0.5 to 4.5. The higher the value, the better the quality of the speech is.

## CHAPTER IV

### Experimental Studies

This chapter investigates the effectiveness and performance of the proposed Speech Enhancement System using Convolutional Denoising Auto-Encoder. The performance of the proposed method has been compared with the performance of Feed-forward Neural Network (FNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) with comparable evaluation metrics. This chapter also provides an experimental analysis for a better understanding of the performance of the proposed method.

#### 4.1 Experimental Setup

The system is implemented in Jupyter Notebook using Keras and Tensorflow in backend. The experiment was performed into a Desktop which has Intel(R) Core™ i7-7770 CPU (@ 4.20 GHz), 16GB of RAM, a Nvidia GTX 1070 8GB GPU. Anaconda distribution was used to incorporate all the python libraries.

#### 4.2 Input Dataset Preparation

The experiment was conducted on Speech Command Dataset [44] which was formed by TensorFlow and Google's AIY team. It was created by following open speech recording from thousands of volunteers over the world. The dataset consists of a vast amount of utterances over 30 specific words which are one second in length. Among them, utterances of numerals were used for this study. The utterances contain ten different sets namely 'zero', 'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight' and 'nine' in total number 24,000 speech signal files. The files were distributed randomly to form the training set and the test set by following the ratio of 70:30. Thus, train set consists of 16,800 utterances and test set consists of 7,200 utterances.

#### 4.3 Noise List

Among the noises described in Section 3.4 [40], 18 types of acoustic noises were considered which can be seen in Table 4.1. They were down-sampled from different ranges to 8 kHz, to properly add them to utterances of the dataset.

**Table 4.1:** List of Noises used for training

Noise Type	Intensity Range (dB)
Speech Babble Noise	40-50
Cockpit Noise	70-80
Engine Room Noise	75-90
Operation Room Noise	40-50
Machine Gun Noise	80-100
Chainsaw Noise	90-110
Fireworks Noise	110-125
Vehicular Interior Noise (Volvo 340)	70-90
Vacuum Cleaner Noise	60-80
Military Vehicle Noise	80-90
Factory Drill Noise	90-110
Factory Floor Noise	80-100
Traffic Noise	50-90
Classroom Noise	40-80
Playground Noise	70-80
White Noise	-
Pink Noise	-
High Frequency Channel noise	-

Intensity of human hearing range varies from 0 dB to 120dB [45] which is the smooth hearing window. After 120 dB, it becomes painful for human ear to perform properly. From table 4.3.1, it is eminent that noise creates a huge distortion in speech if these noises are applied or added to any speech signal.

#### **4.4 MFCC Feature Extraction**

Feature extraction from the utterances is the primary stage for speech enhancement and MFCC features are considered in the proposed system. To extract MFCC features from the audio speech signals, 'librosa' is used in this study [46], which is an open source python library. The library provides the MFCC features of the speech signals by following the same steps discussed in Section 2.4. 16 MFCC filters were applied on each utterance of the clean and noisy data, which generate 16×16 features for each audio signal. The audio signals were

down sampled from 16 kHz to 8 kHz. The voice envelopes which are mimicked by the MFCC filters overlapped with 50ms window which collects the most significant Mel-frequency ceptrums from the speech signals.

#### 4.5 Experimental Results and Analysis

Experimental results of the proposed speech enhancement system have been collected based on the training data which contains 16,800 utterances and test data that consists of 7,200 utterances.

Figure 4.1 depicts the train and test data similarity for the proposed CDAE based speech enhancement system with respect to the number of iterations. Over 1200 iterations were conducted and it can be stated that, the proposed CDAE based speech enhancement system provides 78% test set similarity with only 100 iterations with a batch size of 100. As the number of iterations increased, the model maintained its similarity value. SSIM puts importance on the structural similarity between two 2D objects and calculates the similarity ratio accordingly. CDAE learns potent features with a high rate of abstraction, thus it performs relatively well as we have considered the MFCC features as a 2D structural input to the model. CDAE performs relatively well on the training set to denoise the noisy speech data. The training set similarity went up to 93% where the test set similarity lagged behind a little bit, closing around 80%, with increasing number of iterations. It should be mentioned

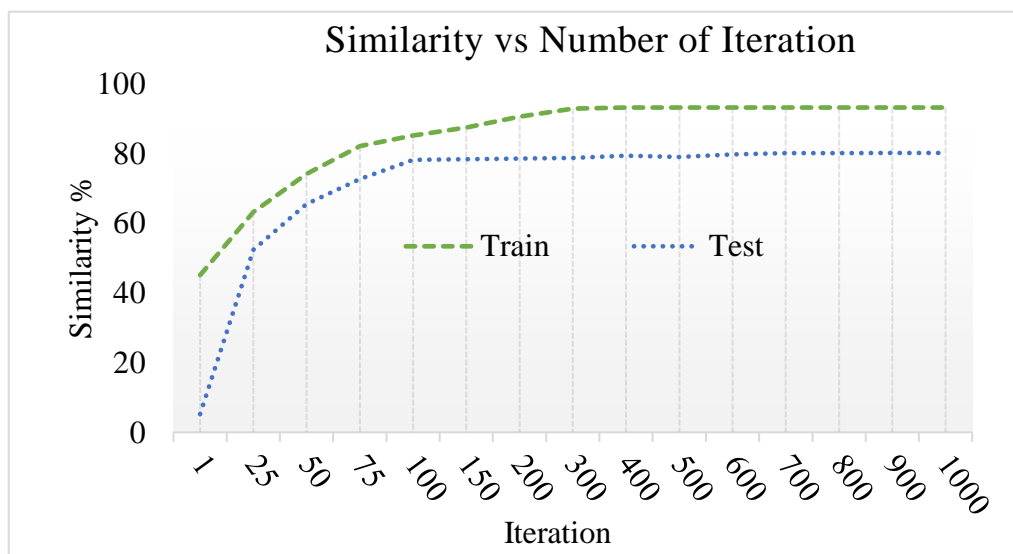


Figure 4.1: Training and Test Sets Similarity for Different Iterations with Batch Size 100.

that higher training set similarity is not as coveted as test set similarity. The figure revealed that the proposed system works significantly well in cleaning noisy speech.

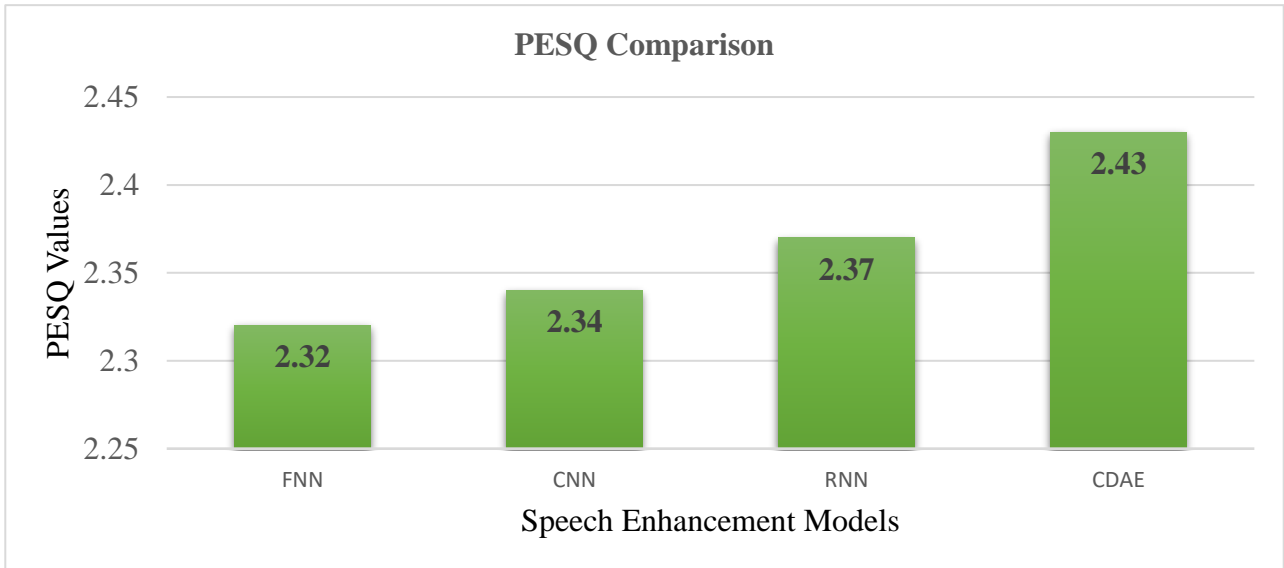


Figure 4.2: Comparison of PESQ Values of Different Speech Enhancement Models.

Figure 4.2 compares achieved Perceptual Evaluation of Speech Quality values by proposed CDAE based system with other models based on FNN, CNN and RNN [5]. FNN gives the lowest PESQ score of 2.32 which is understandable as it only considers the 1D input structure and doesn't consider the local temporal features present in the speech signal. As RNN considers raw speech signals as input, it shows the best PESQ value among the existing methods showing a score of 2.37. On the other hand, performance of CNN is better than FNN with a small margin of 0.02. According to the figure, CDAE performs significantly better than any other models. CDAE gives a PESQ value of 2.43 which states proficiency of the proposed system generating best quality speech from the noisy speech because of its consideration of 2D structured input. Thus, These results prove that CDAE performs better by considering the local temporal features from speech signals.

## CHAPTER V

### Conclusions

Enhancement of speech signals is a tough task which works as a prerequisite for Automatic Speech Recognition (ASR) tasks. Effective Speech enhancement methods are needed to significantly improve the quality and intelligibility of the speech signals. In this thesis a speech enhancement system is proposed using Convolutional Denoising Autoencoder (CDAE). This chapter will draw a short summary of the key points of this thesis and possible future research directions based on the outcome of the present work.

#### 6.1 Contribution

This work introduces a Convolutional Denoising Autoencoder (CDAE) based approach for speech enhancement. CDAE has the ability to recognize spatial relationship between the latent features of the speech signals. The study revealed that 2D representation of the features helped CDAE to learn through convolutional encoder and deconvolutional decoders. The proposed model has been tested on the Speech Command Dataset which is widely available and researched upon for speech related tasks. The proposed method seemed to perform comparatively well with other established models such as FNN, RNN and CNN, in cleaning noisy speeches. The proposed speech enhancement system achieved 80% similarity between the clean speech and denoised speech. The PESQ value of the proposed system is 2.43 which is far better than available prominent systems.

#### 6.2 Future Work

We are looking forward for conducting further study to construct a more general speech enhancement model which considers more unseen noises and further enhance the intelligibility of the speech signals. More evaluation metrics such as Signal-to-Noise Ratio (SNR), Mean opinion score (MOS) etc can be considered in future studies to evaluate the effectiveness of speech enhancement methods more accurately.

## **PUBLICATION FROM THE THESIS**

**Shaikh Akib Shahriyar**, M. A. H. Akhand, N. Siddique and T. Shimamura, "Speech Enhancement Using Convolutional Denoising Autoencoder," *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox'sBazar, Bangladesh, 2019, pp. 1-5.

**DOI: 10.1109/ECACE.2019.8679106**



## REFERENCES

- [1] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, 1979.
- [2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984.
- [3] Y. Ephraim and H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Transactions on Speech and Audio Processing*, 1995.
- [4] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Vol. 2. , 2002.
- [5] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.
- [6] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, 2013.
- [7] H. Sameti, H. Sheikzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, 1998.
- [8] S. W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proceedings - 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017*, 2018.
- [9] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2008.
- [10] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, 2014.

- [11] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2015.
- [12] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2014.
- [13] B. Y. Xia and C. C. Bao, "Speech enhancement with weighted denoising auto-encoder," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013.
- [14] K. Osako, R. Singh, and B. Raj, "Complex recurrent neural networks for denoising speech signals," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2015*, 2015.
- [15] S. W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016.
- [16] T. Yoshioka, S. Karita, and T. Nakatani, "Far-field speech recognition using CNN-DNN-HMM with convolution in time," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2015.
- [17] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1980.
- [18] K. R. Borisagar, R. M. Thanki, B. S. Sedani, K. R. Borisagar, R. M. Thanki, and B. S. Sedani, "Introduction of Adaptive Filters and Noises for Speech," in *Speech Enhancement Techniques for Digital Hearing Aids*, 2018.
- [19] "Multitask Noisy Speech Enhancement System." [Online]. Available: <https://sound.eti.pg.gda.pl/denise/speechband.html>.
- [20] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1979.
- [21] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by

- acoustic noise,” *CASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 4., 1979.
- [22] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: Nonlinear Phenomena*, 1992.
- [23] Y. Xu, J. B. Weaver, D. M. Healy, and J. Lu, “Wavelet Transform Domain Filters: A Spatially Selective Noise Filtration Technique,” *IEEE Transactions on Image Processing*, 1994.
- [24] “The Wiener Filter.” [Online]. Available: [https://dsp-nbsphinx.readthedocs.io/en/nbsphinxexperiment/random\\_signals\\_LTI\\_systems/wiener\\_filter.html](https://dsp-nbsphinx.readthedocs.io/en/nbsphinxexperiment/random_signals_LTI_systems/wiener_filter.html)
- [25] J. Schmidhuber, “Deep Learning in neural networks: An overview,” *Neural Networks*. 2015.
- [26] M. Sun, X. Zhang, H. Van Hamme, and T. F. Zheng, “Unseen noise estimation using separable deep auto encoder for speech enhancement,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2016.
- [27] Y. LeCun, Y. Bengio, and others, “Convolutional networks for images, speech, and time series,” *The handbook of brain theory and neural networks*, 1995.
- [28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, 1998.
- [29] “An intuitive guide to Convolutional Neural Networks.” [Online]. Available: <https://medium.freecodecamp.org/an-intuitive-guide-to-convolutional-neural-networks-260c2de0a050>.
- [30] S. W. Fu, T. Y. Hu, Y. Tsao, and X. Lu, “Complex spectrogram enhancement by convolutional neural network with multi-metrics learning,” in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, 2017.
- [31] T. Kounovsky and J. Malek, “Single channel speech enhancement using convolutional neural network,” in *Proceedings of the 2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics, ECMSM 2017*, 2017.

- [32] A. Pandey and D. Wang, "TCNN: Temporal Convolutional Neural Network for Real-Time Speech Enhancement in The Time Domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.
- [33] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction BT ", *International Conference on Artificial Neural Networks*. Springer, Berlin, Heidelberg, 2011.
- [34] H. L. Y. Bengio, P. Lamblin, D. Popovici, "Greedy Layer-Wise Training of Deep Networks," in *Proceedings of the Advances in Neural Information Processing Systems*, 2007.
- [35] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2015.
- [36] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014.
- [37] "Mel Frequency Cepstral Coefficient (MFCC) tutorial." [Online]. Available: <http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/>.
- [38] "The dummy's guide to MFCC." [Online]. Available: <https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd>.
- [39] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, "Stacked Convolutional Denoising Auto-Encoders for Feature Representation," *IEEE Transactions on Cybernetics*, 2017.
- [40] V. Akkermans *et al.*, "Freesound 2: An Improved Platform for Sharing Audio Clips," *International Society for Music Information Retrieval Conference, Late-breaking Demo Session*, 2011.
- [41] R. Dosselmann and X. D. Yang, "A comprehensive assessment of the structural

- similarity index,” *Signal, Image and Video Processing*, 2011.
- [42] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. Vol. 2., 2002.
- [43] R. ITU-T, “862-perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks,” *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)*, 2001.
- [44] P. Warden, “Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition,” *arXiv preprint arXiv:1804.03209*, 2018.
- [45] S. W. Smith, “The Scientist and Engineer’s Guide to Digital Signal Processing,” in *The Scientist and Engineer’s Guide to Digital Signal Processing*, 1999.
- [46] B. McFee *et al.*, “librosa: Audio and Music Signal Analysis in Python,” in *Proceedings of the 14th Python in Science Conference*, 2018.