

A NOISE FILTERING FRAMEWORK IN MULTI-
CHANNEL SPEECH ENHANCEMENT SYSTEM FOR
ENVIRONMENTAL NOISES

PAVANI CHERUKURU

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITI MALAYA
KUALA LUMPUR

2023

**A NOISE FILTERING FRAMEWORK IN MULTI-
CHANNEL SPEECH ENHANCEMENT SYSTEM FOR
ENVIRONMENTAL NOISES**

PAVANI CHERUKURU

**THESIS SUBMITTED IN FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF DOCTOR OF
PHILOSOPHY**

**FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITI MALAYA
KUALA LUMPUR**

2023

UNIVERSITI MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **PAVANI CHERUKURU** Matric No: **17058250/1/WHA150004**

Name of Degree: **DOCTOR OF PHILOSOPHY**

**A NOISE FILTERING FRAMEWORK IN MULTI-CHANNEL
SPEECH ENHANCEMENT SYSTEM FOR ENVIRONMENTAL NOISES**

Field of Study: **Computer Use**

I do solemnly and sincerely declare that:

- (1) I am the sole author/writer of this Work;
- (2) This Work is original;
- (3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
- (4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
- (5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
- (6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature

Date: 28-1-2023

Subscribed and solemnly declared before,

Witness's Signature

Date: 29-1-2023

Name:

Designation:

A NOISE FILTERING FRAMEWORK IN MULTI-CHANNEL SPEECH ENHANCEMENT SYSTEM FOR ENVIRONMENTAL NOISES

ABSTRACT

The speech enhancement system deals with noisy speech signals by reducing the background noises while preventing any alterations to the speech features. Speech enhancement algorithms are used in multiple channels applied in communication devices to enhance the quality of speech signals under noisy environments known as multi-channel speech enhancement system (MCSE). Micro Electro-Mechanical Systems (MEMS) microphones are used in MCSE systems in outdoor environments. There are many existing algorithms used to filter the noise in speech enhancement systems which are frequently used as a pre-processor to enhance speech quality. These algorithms were effective in the reduction of noisy signals and improved the quality of speech. However, they may have limited ability to perform well on low Signal-to-Noise Ratio (SNR) conditions. The existing MCSE systems can filter 0 to 60dB of SNR, which gives a 62.5% Word Recognition Rate (WRR) at 0dB (considered low SNR), and 83% WRR at 60dB (considered high SNR). However, it was tested only with white Gaussian noise but not with environmental noises, which is very crucial in speech communication devices. Thus, the existing MCSE did not consider all types of noises in a real-time environment. This research aims to propose a noise filtering framework using suitable algorithm(s) for multi-channel speech enhancement systems in handling various Signal-to-Noise ratio (SNR) of environmental noises. This research firstly analyzes the findings of the existing algorithms and components involved in the Speech Enhancement and MCSE systems in handling different types of noises. This is to

identify suitable algorithms for proposing a noise filtering framework for environmental noises. Secondly, experiments were conducted on the existing MCSE as the benchmark systems to analyze the limitations of the existing algorithms in handling environmental noises. From the benchmark experiments, this research has identified that the MCSE's recognition rate reported the highest WRR at 93.77% for high SNR (at 20dB) and 5.64% for low SNR (at -10dB) on an average of five types of different noises. This research has proposed a noise filtering framework that comprises the pre-processing and deep learning algorithms for MCSE in handling various SNRs of environmental noises. The performance of the developed noise filtering framework in handling various SNR of environmental noises shows a WRR of 70.55% at -10dB SNR and 75.44 % at 15dB SNR, while 5.82 % at -10dB and 88.8% at 15dB by the existing MCSE system. It has proven that the proposed pre-processing and deep learning algorithms performed well at low SNR's for MCSE under noisy environments.

Keywords: Multi-Channel Speech Enhancement system, Automatic Speech Recognition System, Speech Enhancement Algorithms, Convolution Neural Network, Bidirectional Long Short Term Memory, Pre-processing algorithm.

RANGKA KERJA PENAPIS BUNYI BAGI SISTEM PENINGKATAN PERTUTURAN BOLEH-DIPAKAI UNTUK BUNYI PERSEKITARAN

ABSTRAK

Sistem peningkatan pertuturan menangani isyarat pertuturan yang bising dengan mengurangkan bunyi latar belakang sambil menghalang sebarang perubahan pada ciri-ciri pertuturan. Algoritma peningkatan pertuturan digunakan dalam pelbagai saluran peranti komunikasi untuk meningkatkan kualiti isyarat pertuturan di bawah persekitaran bising yang dikenali sebagai peningkatan pertuturan berbilang saluran (MCSE). Sistem mikrofon Electro-Mechanical Mikro (MEMS) digunakan dalam sistem MCSE di persekitaran luar. Terdapat banyak algoritma sedia ada yang digunakan untuk menapis bunyi bising dalam sistem peningkatan pertuturan yang kerap digunakan sebagai pra-pemproses untuk meningkatkan kualiti pertuturan. Algoritma ini berkesan untuk pengurangan isyarat bising dan meningkatkan kualiti pertuturan. Walau bagaimanapun, algoritma tersebut mempunyai keupayaan terhad untuk menunjukkan prestasi yang baik pada keadaan Nisbah Isyarat-ke-Bunyi (SNR) yang rendah. Sistem MCSE sedia ada boleh menapis 0 hingga 60dB SNR, yang memberikan Kadar Pengecaman Perkataan (WRR) 62.5% pada 0dB (dianggap SNR rendah), dan WRR 83% pada 60dB (dianggap SNR tinggi). Walau bagaimanapun, ia hanya diuji dengan bunyi *white Gaussian* tetapi tidak dengan bunyi persekitaran, yang mana ia sangat penting dalam peranti boleh pakai. Oleh itu, MCSE sedia ada tidak mengambil kira semua jenis bunyi dalam persekitaran masa nyata. Penyelidikan ini bertujuan untuk mencadangkan

rangka kerja penapisan bunyi menggunakan algoritma yang sesuai untuk sistem peningkatan pertuturan berbilang saluran dalam mengendalikan pelbagai nisbah Isyarat-ke-Bunyi (SNR) untuk bunyi persekitaran. Langkah pertama dalam penyelidikan ini ialah menganalisis penemuan algoritma dan komponen sedia ada yang terlibat dalam sistem Peningkatan Pertuturan dan MCSE dalam menangani pelbagai jenis bunyi. Ini adalah untuk mengenal pasti algoritma yang sesuai untuk dicadangkan sebagai rangka kerja penapisan bunyi untuk bunyi persekitaran. Kedua, eksperimen dijalankan ke atas MCSE sedia ada sebagai sistem penanda aras untuk menganalisis batasan algoritma sedia ada dalam menangani bunyi persekitaran. Dari eksperimen penanda aras, penyelidikan ini telah mengenal pasti bahawa kadar pengecaman MCSE melaporkan WRR tertinggi pada 93.77% untuk SNR tinggi (pada 20dB) dan 5.64% untuk SNR rendah (pada -10dB) secara purata bagi lima jenis bunyi yang berbeza. Penyelidikan ini telah mencadangkan rangka kerja penapisan bunyi yang terdiri daripada algoritma pra-pemprosesan dan pembelajaran mendalam untuk MCSE dalam mengendalikan pelbagai SNR bunyi persekitaran. Prestasi rangka kerja penapisan bunyi yang dibangunkan dalam mengendalikan pelbagai SNR bunyi persekitaran menunjukkan WRR sebanyak 70.55% pada -10dB SNR dan 75.44% pada SNR 15dB, manakala 5.82% pada -10dB dan 88.8% pada 15dB oleh sistem MCSE sedia ada. Ia membuktikan bahawa algoritma pra-pemprosesan dan pembelajaran mendalam yang dicadangkan menunjukkan prestasi yang baik pada SNR rendah untuk MCSE dalam persekitaran yang bising.

Kata kunci: : Sistem Peningkatan Pertuturan Pelbagai Saluran, Sistem Pengecaman Pertuturan Automatik, Algoritma Peningkatan Pertuturan, Rangkaian Neural Berlingkaran, Memori Jangka Pendek Panjang Dua Arah, Teknik Pra-pemprosesan

ACKNOWLEDGEMENTS

First and foremost, I am thankful to God. He has given me strength and encouragement throughout all the challenging moments of completing this dissertation. I am truly grateful for his unconditional and endless love, mercy, and grace.

Consequently, I would like to express my deepest appreciation to my supervisor **Associate Professor Dr. Mumtaz Begum Mustafa**. She always provided her heartfelt guidance and have given me invaluable guidance, inspiration, and suggestion in my quest for knowledge. She has given me all the freedom to pursue my research, while silently and non-obtrusively ensuring that I stay on course and do not deviate from the core of my research. Without guidance and persistent help of hers, this thesis would not have been possible.

Most important, I would like to express my gratitude to my family and parents for being a source of support and encouragement. My father Bhaskar Naidu C and mother Lalithamma C who were always there to shower me with all-round love and support. Special appreciation goes to my husband Madhusudan G, who has always been with me, he has always pushed and motivated me when I was down. He persistently supported me, incorporate and encouraged my path towards achieving my Ph.D. Moreover, I am grateful to my lovely kid Maanvitha for her innocent smile, may god save and bless her.

My deepest appreciation with gratitude goes to my sisters Jyothsna and Divya. I am really proud of them. They both supported and encouraged me on achieving my Ph.D.

Finally, special thanks to my fellow friends in the Human-Computer Interaction (HCI) Lab, University of Malaya as well as those who were involved directly or indirectly towards completion of this thesis.

TABLE OF CONTENTS

Abstract	iii
Abstrak	v
Acknowledgements	vii
Table of Contents	viii
List of Figures	xv
List of Tables	xvii
List of Symbols and Abbreviations	xix
List of Appendices	xxiii
CHAPTER 1: INTRODUCTION	26
1.1 Speech and Noise	26
1.2 Research Background	30
1.2.1 Speech Enhancement System	30
1.3 Research Motivation	36
1.4 Problem Statement	37
1.5 Research Objectives	40
1.6 Research Questions	40
1.7 Research scope	41
1.8 Research Methodology	41
1.9 Thesis Organization	43
CHAPTER 2: LITERATURE REVIEW	45

2.1	Overview of this Chapter	45
2.2	Noise in Speech Signals	45
2.2.1	Speech sources	46
2.2.1.1	Non-stationarity.....	46
2.2.1.2	Wideband signal.....	46
2.2.1.3	Non-Gaussianity.....	46
2.2.1.4	Human Speech production	47
2.2.1.5	Pitch.....	47
2.2.2	Music sources.....	48
2.2.3	Noise sources.....	48
2.3	Speech enhancement.....	51
2.3.1	Multichannel Speech Enhancement.....	51
2.3.1.1	Multi-sensory Beamforming Algorithms used for speech enhancement.....	53
2.3.1.2	Maximum Signal-to-Noise Ratio (Max-SNR) Beamformer.....	57
2.3.1.3	Delay and Sum Beamformer (DSB).....	57
2.3.1.4	Adaptive Noise Reduction/ Cancellation Algorithms.....	58
2.3.2	Deep Learning-Based Algorithms On Speech Enhancement System...	60
2.3.3	Summary of Speech enhancement Algorithms.....	68
2.4	Multi-Channel Speech Enhancement	69
2.4.1.	The performance of the Multi-Channel Speech Enhancement systems in the existing studies.....	69

2.4.2. Summary of Multi-Channel Speech Enhancement.....	75
2.5 Speech Enhancement for Automatic Speech Recognition.....	75
2.5.1 Preprocessing Algorithm.....	75
2.5.2 Speech Classification.....	81
2.6 Theoretical Background and Implementation Requirements.....	83
2.6.1 Discrete Wavelet Transform.....	83
2.6.2 Convolution Neural Network.....	86
2.7 Performance Evaluation Methods for Speech Enhancement Algorithms.....	94
2.8 Summary.....	98
CHAPTER 3: THE MULTI CHANEL SPEECH ENHANCEMENT SYSTEMS IN NOISY ENVIRONMENT: THE BENCHMARK EXPERIMENT.....	99
3.1 Overview.....	99
3.2 Problem Identification and the proposed solution.....	99
3.3 Dataset	100
3.4 Benchmark Experiment: Multi-Channel Speech Enhancement (MCSE).....	101
3.4.1 Beamforming.....	103
3.4.2 Adaptive Noise Reduction (ANR).....	104
3.4.3. Voice Activity Detection (VAD).....	105

3.5	Experimental Design and Setup	107
3.5.1	Experimental design for Multi-Channel Speech Enhancement.....	107
3.5.2	Experimental Setup for Multi-Channel Speech Enhancement.....	108
3.6	Evaluation.....	109
3.6.1	Spectrogram Analysis.....	109
3.6.2	Word Error Rate (WER).....	109
3.7	Results.....	110
3.7.1	Spectrogram Analysis.....	109
3.7.2	Word Error Rate (WER).....	119
3.8	Discussion.....	124
3.9	Summary.....	125
CHAPTER 4: THE PROPOSED NOISE FILTERING FRAMEOWRK FOR MULTI-CHANNEL SPEECH ENHANCEMENT SYSTEM.....		126
4.1	Overview.....	126
4.2	The Proposed Framework for Multi-Channel Speech Enhancement System... 	126
4.2.1.	Discrete Wavelet transform.....	128
4.2.2	Deep Learning-based approach.....	130
4.3	Dataset details.....	140
4.4	Experimental Design and Setup.....	142
4.4.1	Experimental Design.....	142
4.4.2.	Experimental Setup.....	143

4.4.2.1 Sampling Setup	
4.4.2.2. Variability Setup	
4.4.2.3 Noise and Sample Utterance System Setup	
4.4.2.4 SNR Setup	
4.5. Software Requirements.....	144
4.6 Performance measurement parameters.....	147
4.7 The differences between existing MCSE and the proposed deep learning based noise filtering framework.....	150
4.8 Summary.....	151
CHAPTER 5: EVALUATION, RESULTS AND DISCUSSION.....	152
5.1 Overview.....	152
5.2 Evaluation.....	152
5.2.1 Spectrogram Analysis.....	152
5.2.2 Word Error Rate (WER).....	153
5.2.3. Performance Measurement Parameters.....	153
5.3 Results.....	155
5.3.1. Spectrogram Analysis for Multi-Channel Speech Enhancement.....	155
5.3.2. Word recognition rate (WRR) for the proposed Deep learning based Multi-Channel Speech Enhancement system.....	163
5.3.3 Performance measurement parameters.....	166
5.4 Performance comparison with existing methods.....	18569

5.5 Discussion	18572
5.6 Summary.....	178
CHAPTER 6: CONCLUSION AND FUTURE.....	179
6.1 Overview.....	179
6.2 Fulfilment of Research Objectives	179
6.2.1 Research Objective 1.....	179
6.2.2 Research Objective 2.....	181
6.2.3. ResearchObjective 3.....	182
6.2.4. Research Objective 4.....	183
6.3 Research Contributions.....	185
6.4 Research Limitation.....	186
6.5 Suggestions for Future Works	186
References.....	188
List of Publications and Papers Presented.....	210
Appendices.....	211
Appendix A.....	211
Appendix B	212
Appendix C	213
Appendix D	217

LIST OF FIGURES

Figure 1.1: Sources of Noise in Speech Communication.....	31
Figure 1.3: Research methodology.....	34
Figure 2.1: Speech production model (Edmund Lai et al., 2003).....	51
Figure 2.2: Spectrogram of speech signal (a) clean signal (b) babble noise (c) train noise and (d) Gaussian noise (Obtained from the Matlab simulation code).....	52
Figure 2.3: Generalised Architecture of Multi-channel Speech enhancement.....	55
Figure 2.4: Delay and sum beamformer with J microphones.....	62
Figure 2.5: Generalized Architecture of Deep learning based Speech Enhancement.....	67
Figure 2.6: Architecture of CNN in speech enhancement (Se Rim et al., 2016).....	75
Figure 2.7: Generalised Architecture of Multi-Channel Speech Enhancement.....	91
Figure 2.8: Wavelet transform (Hazrat et al., 2014).....	102
Figure 3.1: The architecture of Multi-Channel Speech Enhancement (MCSE)	114
Figure 3.2: Real time architecture of Multi-Channel Speech Enhancement embedded in Helmet.....	115
Figure 3.3: Clean speech signal.....	123
Figure 3.4: The spectrogram for unfiltered white Gaussian noise in speech signal at -5db...	125
Figure 3.5: The filtered white Gaussian noise in speech signal at -5db.....	126
Figure 3.6: The unfiltered Airport noise in speech signal at -5dB.....	126
Figure 3.7: The filtered Airport noise in speech signal at -5dB.....	126

Figure 3.8: The unfiltered Babble noise in speech signal at -5dB.....	127
Figure 3.9: The filtered Babble noise in speech signal at -5dB.....	127
Figure 3.10: The unfiltered Car noise in speech signal at -5dB.....	128
Figure 3.11: The filtered Car noise in speech signal at -5dB.....	128
Figure 3.12: The unfiltered exhibition noise in speech signal at -5db.....	129
Figure 3.13: The filtered exhibition noise in speech signal at -5db.....	129
Figure 3.14: The unfiltered restaurant noise in speech signal at -5db.....	130
Figure 3.15: The spectrogram for filtered restaurant noise in speech signal at -5db.....	130
Figure 3.16: The SNR and WRR linear relationship for stationary noise in MCSE.....	132
Figure 3.17: The SNR and WRR linear relationship for non-stationary noise in MCSE....	134
Figure 3.18: WRR for both the stationary and non-stationary noises in MCSE.....	135
Figure 3.19: Result of ANOVA for stationary and non stationary in MCSE.....	136
Figure 4.1: The Proposed Multi-Channel Speech Enhancement Framework based on Wavelet transform and Deep learning approach (CNN-BLSTM).	139
Figure 4.2: CNN-BLSTM Architecture applied for speech enhancement.....	144

Figure 4.3: An example of CNN process (Dong wang et al., 2019).....	148
Figure 4.4: LSTM unit (Dong wang et al., 2019)	151
Figure 4.5: BiLSTM architecture (Dong wang et al., 2019).....	152
Figure 5.1: Clean speech signal	162
Figure 5.2: WRR for both stationary and non-stationary noise.....	187
Figure 5.3: Results of ANOVA.....	187
Figure 5.4: WRR of Existing Vs Proposed MCSE under non stationary environment...	189
Figure 5.5: Results of ANOVA.....	190
Figure 5.6: WRR of Existing Vs Proposed MCSE under non stationary environment..	190
Figure 5.7: Results of ANOVA	

LIST OF TABLES

Table 1.1 Noises with SNR levels considered in existing papers.....	31
Table 1.2 Comparison between single and multi-channel enhancement with the identified factors.....	36
Table 1.3 Comparison of single channel enhancement and multi-channel enhancement with their methods and filters.....	36
Table 2.1: Existing studies on Deep learning algorithms along with their metrics, data base used, results, advantages and disadvantages.....	67
Table 2.2: Multi-Channel Speech enhancement algorithms using MEMS microphones.....	71
Table 2.3: Comparison of preprocessing algorithms based on various measuring factors.....	79
Table 3.1: Dataset used for experimenting Multi-Channel Speech Enhancement....	103
Table 3.2: Experimenting the Multi-Channel Speech Enhancement (MCSE) includes Beamforming, Adaptive noise reduction and Voice activity detection algorithms Using Different Types of Environmental Noises at Different SNR Levels	109
Table 3.3: Word Recognition Rate (WRR) For Multi-Channel Speech Enhancement (under Stationary Environmental Noise	121
Table 3.4: Word Recognition Rate (WRR) for Multi-Channel Speech Enhancement (MCSE) under Non-stationary Environmental Noises.....	123
Table 4.1: Detail Explanation of each Component and their process in proposed approach.	129
Table 4.2: Detailed explanation of each layer of CNN with their function, libraries and their work process.....	140
Table 4.3: Experimental design of proposed noise filtering framework (Deep learning based) in Multi-Channel Speech Enhancement	143
Table 5.1: Description of the parameters for performance evaluation.....	156
Table 5.2: Spectrogram analysis of Airport noisy and enhanced speech signal at different SNR's using proposed Deep learning approach.....	158

Table 5.3: Spectrogram analysis of Babble noisy and enhanced speech signal at different SNR's using proposed Deep learning approach.....	160
Table 5.4: Spectrogram analysis of Restaurant noisy and enhanced speech signal at different SNR's using the proposed Deep learning approach.....	161
Table 5.5: Spectrogram analysis of Additive white gaussian noise (AWGN-non stationary) noisy and enhanced speech signal at different SNR's using the proposed Deep learning approach	163
Table 5.6: WRR and WER performance by using proposed Deep learning approach	165
Table 5.7: Average results of WRR performance on both stationary and non-stationary by using proposed Deep learning approach.	167
Table 5.8: performance analysis for airport noise using proposed Deep learning approach based MCSE.....	168
Table 5.9: Performance analysis for Babble noise using proposed Deep learning based MCSE.....	168
Table 5.10: Performance analysis for Restaurant noise using proposed Deep learning based MCSE.....	169
Table 5.11: Performance analysis for AWGN noise using proposed Deep learning based MCSE.....	170
Table 5.12: Comparison of proposed MCSE with Existing MCSE system Under Stationary noises at different levels of SNR.....	170
Table 5.13: Comparison of proposed MCSE with Existing MCSE system Under Stationary noises at different levels of SNR.....	172

LIST OF SYMBOLS AND ABBREVIATIONS

ASR	:	Automatic Speech Recognition
SNR	:	Signal-to-Noise ratio
Db	:	Decibels
AWGN	:	Additive White Gaussian Noise
LMS	:	Least Mean Square
ANR	:	Adaptive Noise Reduction
VAD	:	Voice Activity Detection
DMA	:	Directional Microphone Array
SS	:	Spectral Subtraction
NSS	:	Non-linear spectral subtraction
NWNS	:	Non-linear Weighted Noise Subtraction
MCSE	:	Multi-Channel Speech Enhancement
PESQ	:	Perceptual Evaluation of Speech Quality
VDCNN	:	Very deep convolution neural network
WRR	:	Word Recognition Rate
RF	:	Radio Frequency
ANC	:	Adaptive Noise Cancellation
MBBS	:	Multiband Spectral Subtraction
MFCC	:	Mel-Frequency Cepstral Coefficients
PCA	:	Principal Component Analysis
LPCC	:	Linear Predictive Cepstral Coefficient
DWT	:	Discrete Wavelet Transform

WPT	:	Wavelet Packet Transform
PDF	:	Probability density function
AR	:	Autoregressive
VOIP	:	voice over internet protocol
FIR	:	finite impulse response
NLMS	:	Normalized Least Mean Square
SD	:	Speech Distortion
DSB	:	Delay and Sum Beamformer
RLS	:	Recursive least squares
MVDR	:	minimal variance distortion less-response
DNN	:	Deep neural networks
TF	:	Time-frequency
CM	:	complex masks
MLP	:	multilayer perception
AdMBSS	:	adaptive multi-band spectral subtraction
MOS	:	mean opinion score
BAK	:	intrusiveness of background noise
SIG	:	signal distortion
WA	:	word accuracy
SegSNR	:	segmental signal-to-noise ratio
SDR	:	source-to-distortion ratio
STOI	:	short-time objective intelligibility
DAE-MFCC	:	Deep autoencoder based on MFCC

RNN-LSTM	:	Recurrent neural network-Long short term memory
CNN	:	Convolution neural network
GAN	:	generative adversarial network
ANN	:	Artificial neural network
ReLU	:	Mel Frequency Cepstral Coefficients
PReLU	:	proportional rectified linear units
CE	:	Cross Entropy
HF	:	Hessian-free
FCN	:	Fully connected layer
BLE	:	Bluetooth Low Energy
PD	:	Parkinson's disease
ARQ	:	automated repeat request
TDS	:	Tongue-Drive System
MEMS	:	Micro-Electro-Mechanical Systems
ECMs	:	Electret Condenser Microphones
GUI	:	graphical user interface
PD	:	Parkinson disease
RNN	:	Recurrent Neural Network
OQCM	:	overall quality composite measure
WSM	:	weighted scoring method
VoIP	:	Voice over Internet Protocol
PD	:	Partial Discharges
WER	:	Word Error Rate

MSE	:	mean square error
IS	:	Itakura-Saito
CC	:	Cepstrum coefficients
LPC	:	Linear Predictive Coding
STOI	:	Short time objective intelligibility
STT	:	speech to text engine
ZCR	:	Zero Crossing Rate
ADC	:	Analog to Digital converter
SD	:	Standard Deviation
RAM	:	Random Access Memory
SPL	:	Sound Pressure levels
CWT	:	Continuous Wavelet Transform
BLSTM	:	Bidirectional Long Short-Term Memory
BN	:	Batch-normalization
TDT	:	Tucker Davis Technology
DAQ	:	Data Acquisition
DSP	:	Digital Signal Processors
FIR	:	Finite Impulse Response
IIR	:	Infinite Impulse Response
Covl	:	overall speech quality
Cbak	:	compound estimate for noise distortion
LH	:	Low-High
HL	:	High-Low

HH : High-High
LL : Low-Low
CEP : Cepstrum Distant Measures

Universiti Malaya

LIST OF APPENDICES

Appendix A: Noise Maker to create -10db noisy speech signal	223
Appendix B: AURORA Noisy speech signals from and noisy speech signals recorded by MEMS microphones	224
Appendix C: Benchmark experiments with beamforming, ANR and VAD algorithms Beamforming Source code	225
Appendix D: The Proposed noise filtering framework using DWT and CNN-BLSTM Discrete Wavelet Transform	254

CHAPTER 1: INTRODUCTION

1.1 Speech and Noises

The primary mode of communication for all human beings is the speech. Speech is a sequence of sounds and Sound is produced by altering the flow of air from the lungs to the mouth through various tuners i.e., tongue, chin, lips, etc. One of our most basic needs is to be able to express ourselves verbally. Speech is the most effective and cheap mode of communication for a wide range of reasons. A speaker's mood can be conveyed through their speech as well as their linguistic content. Communication is often easier and more precise when speakers and listeners are close to each other in a quiet atmosphere. The listener's capacity to understand is hampered when the speaker is far away, or the environment is noisy. When it comes to efficient data sharing, clear and understandable speech is critical in many speech-based systems. As a result, in real life, the existence of additive background and channel noise severely affects the effectiveness of speech processing systems, resulting in erroneous information exchange and weariness among listeners. Several algorithms for improving speech quality from damaged speech have been developed over the years by scholars across the globe. While this may sound like a little problem, it is in fact a very difficult one to solve in the field of speech processing and communication systems research.

Speech processing is a branch of science that examines how various signal processing algorithms might be used to reduce the amount of noise in damaged speech. Speech processing, in general, entails acquiring the speech signal, processing it, storing it for later use, transmitting it to the desired location, and generating an output. The input of speech processing is the speech recognition and the output is called speech synthesis. The application of speech processing is the goal of speech enhancement to improve human perception or computer decoding of loud voice signals. An attempt is made by speech enhancement algorithms to boost the effectiveness of

communication systems when their input or output signals have been distorted by noise. The quality and comprehension of speech degrades when there is a lot of background noise. Naturalness and recognizability are examples of qualities that are included in speech quality. What the speaker actually said, in other words, the meaning or information content of what they were saying, is the focus of intelligence. As a result, the capacity of the speaker and listener to communicate is compromised in a noisy situation. Speech enhancement can be used to lessen the impact of this issue. Trade-offs between reducing noise and improving voice quality limit the ability of speech enhancement systems to perform at their full potential.

Speech application and speech recognition, hearing devices, speech communication systems, and other speech applications can all benefit from efforts to improve the quality and/or intelligibility of loud speech (Adeel et al., 2020). According to specific applications, the purpose of speech enhancement can be to reduce listener fatigue, promote overall speech quality, increase intelligibility, and to better speech communication devices (Darabkh et al., 2018). In order to avoid the loss of speech quality and to overcome the shortcomings of human auditory systems, speech enhancement is necessary.

Speech communication takes place in a variety of settings, including the workplace, doctor's office, and school (Donahue et al., 2018). Communication is extremely secure and exact when both the speaker and the listener are in close proximity to each other in a silent medium. Random external noises tend to affect the quality of verbal communication between them when they are a long distance away. Despite the fact that noises can be found everywhere, it is impossible to identify them all (Vincent et al., 2018).

Noises may have well-known or unknown characteristics, but they all have the potential to interrupt, distort, or degrade speech transmissions. Thus, people with hearing loss may be affected by noise in the environment. In speech recognition, noisy environments are classified as stationary and non-stationary noisy environment. In our everyday surroundings, we frequently hear non-stationary background noises, which refers to the background noises we hear during real conversations. Consequently, stationary noises are synonymous to the noise on telephone lines (Leman et al., 2018). Usually, in our everyday environment, we hear temporary background noise related to the background noise we hear in real conversations. Steady noise, on the other hand, is similar to telephone line noise (Leman et al., 2008). It is a challenge to remove background noise, such as fan noise, car noise, and other intervening speakers, from a speaker's speech spectrum in order to produce speech signals with a better perceived quality.

The use of speech enhancement in a noisy environment is common. Speech enhancement algorithms can be used to reduce the amount of noise in a noisy speech signal and enhance its clarity. There are a variety of algorithms for enhancing speech in both single and multi-channel environments (Gabbay et al., 2018; Novotny et al., 2019).

Figure 1.1 depicts a block diagram of various noise sources and it shows the noise speech signal, which is a blend of background noise, intended speaker, communication noise, and other speaker noise. It is via the use of speech enhancement algorithms that the original noisy speech signal can be transformed into a clear, easily audible signal (Vincent et al., 2018).

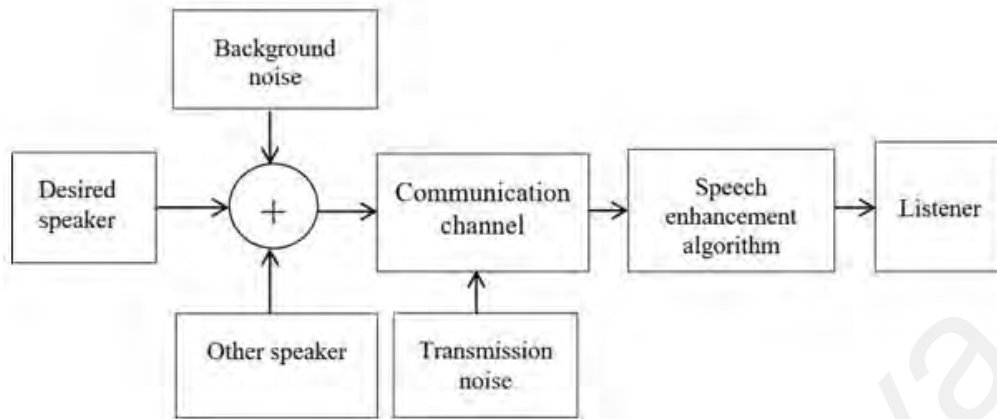


Figure 1.1: Sources of Noise in Speech Communication

Disruptions can make a discussion awkward in the worst-case scenario, depending on the Signal-to-Noise Ratio (SNR). SNR's may vary from low to high decibels(dB) in any kind of speech to noisy signals i.e., -10dB, -5dB, 0dB, 5dB, 10dB, 15dB, 20dB etc. The SNR power ratio is expressed in decibels (Pinki et al., 2015). It is the most widely accepted and well-liked method for evaluating speech quality. Table 1.1 shows the example on SNR levels with noises considered in existing literature.

Table 1.1: Noises with SNR levels considered in existing papers

Noises	SNR'S	References
White Guassian Noise	0dB, 5Db, 10Db, 15dB SNR's considered for evaluation	Shanmugapriya et al., 2014
Train, Car, Babble noise	5db train, 10db car, 5db babble noise	Yi Hu et al., 2007
AWGN, Exhibition, Station, Drone, Helicopter, Airplane	0dB, 2.5dB, 5Db	Rahul Kumar Jaiswal et al., 2022

Speech enhancement algorithms can help reduce background sounds and echoes without damaging the speech stream through digital signal processing in order to deal with these kinds of

acoustic situations (Taha et al., 2018). Key to the success of the noise reduction system is to improve SNR, clarity, and computational complexity. The intelligibility and quality of speech cannot be improved simultaneously by any speech enhancement system. If a person's speech is easily understood, it is likely to be deemed high-quality; conversely, a person's speech that cannot be understood should be considered low-quality.

1.2 Research Background

1.2.1 Speech Enhancement System

Speech enhancement processes noisy speech signals by reducing background noise while preventing changes in speech characteristics. Speech enhancement is used in voice signal processing applications such as voice coders, automatic speech recognition, Voice over Internet Protocol, and hearing aids.

Single microphones and enhancing algorithms are used in speech enhancements system, whereas MEMS microphones (multi) are used in Multi-Channel Speech Enhancement system. There are many existing algorithms used to filter noise in speech enhancement systems and are often used as a preprocessor to improve speech quality. They have proven effective in reducing interference signals and improving voice quality. However, under low signal-to-noise ratio (SNR) conditions, their ability to achieve well may be limited.

The single channel is typically not available in most real-time applications such as speakers, voice recognition, mobile communications, and hearing aids. These systems are easy to build and relatively cheaper than multi-input systems. This is one of the most difficult situations in speech enhancement because there is no reference signal available for noise and clean speech/audio signal cannot be preprocessed before it is affected by the noise (Yadava et al., 2019).

Approaches for enhancing speech with only one acquisition channel are known as "single-channel" algorithms. An application's ability to use certain signals may be constrained by the system it uses (such as with telephone-based applications and pre-recorded applications) (Clark, et al., 2019). Spectral subtraction (SS) is a direct method of improving noisy speech when the noise process is stationary and speech activity can be recognized (Vinay et al., 2021). Single-channel systems typically use different voice and unwanted noise statistics. Most algorithms assume that the noise is stationary during the audio interval, so the performance of these algorithms is usually limited to the presence of non-stationary noise. Also, low signal-to-noise ratios can significantly reduce performance (Yan et al., 2020).

The multi channel systems take advantage of the availability of multiple signal inputs to the system, use noise references in adaptive noise cancellers, use phase adjustment to cancel unwanted noise components, and combine step-by-step schemes (Kokkinakis and Loizou, 2010). By considering the spatial characteristics of the signal and noise source, the limitations inherent in single-channel systems, especially transient noise, can be adequately addressed in the multi-channel systems. These systems are usually more complex.

Figure 1.2 depicts the architecture of multi-channel speech enhancement. The architecture consists of Microphone array, Beamforming, Adaptive noise reduction and Voice activity detection.

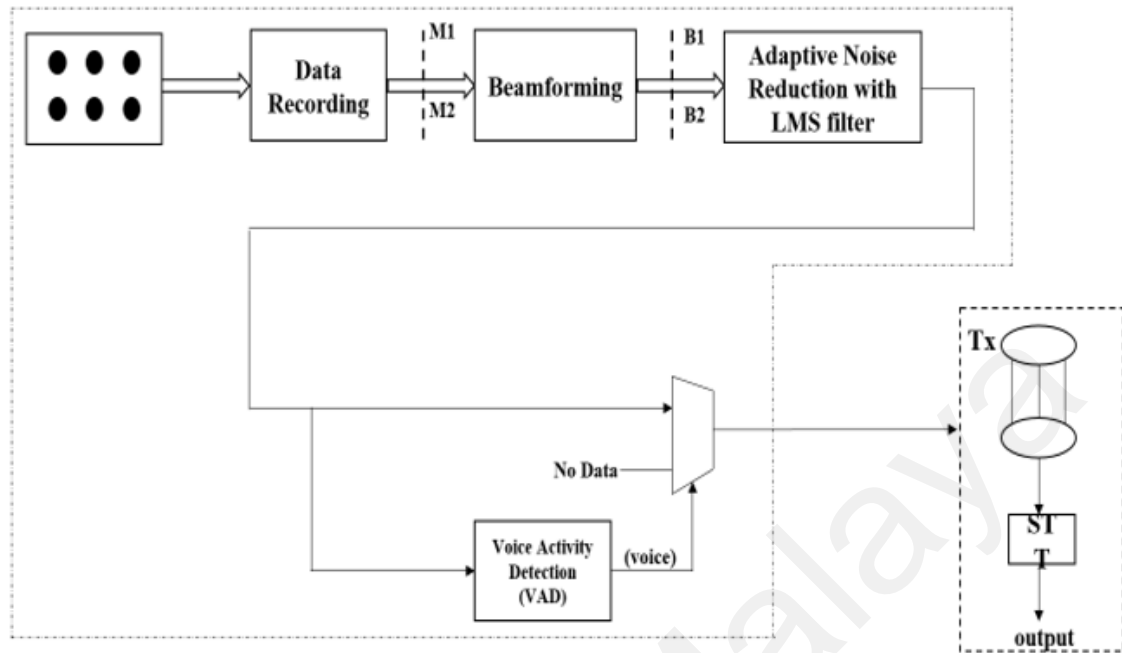


Figure 1.2: Architecture of Multi-channel Speech Enhancement (MCSE) (Alessandro et al., 2017)

- Beamformer is a signal processor used in combination with a microphone array to provide spatial filtering capabilities. Beamforming can be achieved by filtering the microphone signal, combining the outputs to extract the desired signal, and eliminating interference noisy signals (Van Veen et al., 1988). There are two types of beamforming, fixed beamforming and adaptive beamforming. In fixed beamforming where the direction of the input signal is fixed, the distance between the microphones is constant. Fixed beamforming can be achieved using delay and sum beamformers. Adaptive beamforming is one in which the directivity of the input noisy speech signal changes in response to changes in the acoustic environment (Ramesh Babu et al., 2015).
- Adaptive Noise Reduction (ANR) is used to filter environment noise with Least Mean Square (LMS) filter (Soo et al., 1990; Valin et al., 2007). The user beam and reference noise ($b_1 \dots b_n$) are the inputs to the ANR. The ANR component filters the noise of the user

beam interconnected with the reference noise, but the audio signal is only present in the user beam as already processed with beamforming, but is not attenuated (Widrow et al., 1975).

- Voice Activity Detection algorithm is to distinct the presence of the user's voice in the user stream (Venkatesha Prasad et al., 2002). This could be useful for two reasons:
 - (i) Segmentation: the system needs to know the exact boundaries of each word in spoken utterance.
 - (ii) Data Reduction: the system only send data as required and not sending the data continuously over the transmission channel.

The problem faced by speech enhancement algorithms is to estimate the voice signal from a corrupted version of the signal that consists of the desired voice and noise. The complexity of the problem is relatively high due to the limited information available in a single observation (Bachu et al., 2008).

Existing speech enhancement algorithms such as spectral subtraction, non-linear spectral subtraction utilize a single microphone to process audio signals (Ganga Prasad et al., 2013). However, these algorithms are computationally intensive and are not effective at suppressing noisy audio signals, especially when the SNR is low i.e., -10 dB to 10 dB (Ken Chen et al., 2012). This noise is difficult to filter because it has different characteristics in terms of noisy levels in decibels, frequencies etc., depending on the environment. Therefore, multi-channel speech enhancement is very much required (Mohammed Akhaee et al., 2005). The comparison of single and multi-channel speech enhancement is presented in Table 1.2 and Table 1.3 respectively.

Table 1.2: Comparison between single and multi-channel enhancement with the identified Factors

Factors	Speech Enhancement Algorithms	
	Single Channel Enhancement	Multi-Channel Enhancement
Recording	Uses only one microphone for recording	Uses more than one microphone for recording
Filtering	Filters only stationary noise	Non-stationary noise is better handled by these systems as compared to single channel.
Implementation	This system is easier to implement, and cost is effective	More complex to implement and very expensive
Sources of input	Single signal input due to single channel	The numerous signal inputs that are available to the system are utilised by the adaptive noise cancelling mechanism in this system.
Estimation of noise signal	Very difficult to estimate noise	This can do better to estimate noise than single channel
Performance	Performance is limited in the presence of non-stationary and degraded with lower SNR	-
Reduction of noise	It is only for reducing background noise	It is not only for background noise but also to reduce the effects due to reverberation and other weak interfering signals.

Table 1.3: Comparison of single channel enhancement and multi-channel enhancement with their methods and filters

References	Algorithm	Methods	Algorithms / Filters	Advantages	Disadvantages
Shanmugapriya et al., (2014), Upadhyay et al., (2015)	Single Channel enhancement	Spectral subtraction method	Weiner Filtering	<ul style="list-style-type: none"> • very simple and easy for implementation 	<ul style="list-style-type: none"> • Musical noise is introduced into the analysis via this approach, due to a mismatch between the estimated and actual noise. • Slightly improved in SNR ratio
			Kalman filtering		

		Over subtraction method	-	<ul style="list-style-type: none"> • very simple and easy for implementation 	<ul style="list-style-type: none"> • Musical noise is a major limitation
		Non-linear spectral subtraction (NSS)	-	<ul style="list-style-type: none"> • very simple and easy for implementation • Removes musical noise 	<ul style="list-style-type: none"> • Real speech is removed.
		Non-linear Weighted Noise Subtraction (NWNS)	-	<ul style="list-style-type: none"> • Improved the performance of voice in noisy environments. 	<ul style="list-style-type: none"> • complete noise cancellation is infeasible
		Multiband Spectral Subtraction (MBSS)	-	<ul style="list-style-type: none"> • Effectively reduces residual noise tones and enhances overall voice quality at low SNRs. 	<ul style="list-style-type: none"> • complete noise cancellation is infeasible
Anand Krishna et al., (2016)	Multi-Channel enhancement	Adaptive Noise Cancellation	LMS	<ul style="list-style-type: none"> • Reference noise can be generated. • Easy to estimate the noise 	<ul style="list-style-type: none"> • More complex to build. • Poor performance at high noisy signals • No improvement at very low SNR's
			NLMS		
RLS					
Weiner filtering					
			Kalman filtering		
		Beamforming	-	<ul style="list-style-type: none"> • Easy to implement. <ul style="list-style-type: none"> i. 	<ul style="list-style-type: none"> • Poor performance at high noisy signals • A lot of sensors are needed to increase the signal-to-noise ratio. • No improvement at very low SNR's

In single channel enhancement, there are many algorithms experimented such as subtraction method, over subtraction method, non-linear spectral subtraction, non-linear weighted subtraction etc. These are very simple and easy for implementation in computing. These algorithms improved the performance of speech quality in noisy environments, but complete noise cancellation is infeasible (Shanmugapriya et al., 2014; Upadhyay et al., 2015).

In multi-channel enhancement, multi sensor beamforming and adaptive noise reduction algorithms were experimented. These are easy to implement but more complex to build. All these algorithms are good in recognition rate but fail to perform better at high noisy environments (Anand Krishna et al., 2016).

1.3 Research Motivation

Currently, the digital communication is widely adopted where various types of communication devices are used. With the help of these devices, information can be exchanged to a remote location such as telephone, television, Bluetooth devices and RF transceivers. During the communication, the data is exchanged over a wireless medium which suffers from various contaminations such as interference noises that are caused from various sources. Thus, maintaining low noise or mitigating the noise has remained a primary challenging task for the research community. In this field of speech enhancement, several algorithms have been introduced to deal with the noise related issues. There are issues in improving performance at low levels of SNR's (-10db, 0db, 5db) which is yet to be solved in this research area.

Due to the current technological advancements, the research community has developed various speech communication devices such as wearable watch, speech translator, speech devices for people hearing impairments. Wearable hearing aids, wearable microphones, wearable watches,

wearable system without hermetic packing are the devices that use multi-channel speech enhancement at present (Seon Man Kin et al., 2020). These devices have been proved as a significant breakthrough to facilitate communication and it can be an assistive tool for people suffering from various types of disabilities such as hearing and motor impairments. It can assist children with autism in communication and impaired people with physical weakness to communicate with computers through their voice. Moreover, this technology can be used to enhance the livelihood of impaired children by providing them new ways to learn various things in an understandable, enjoyable, and desirable manner. However, they perform better at high SNR's but not at low SNR's. There is a lack of research implementing AI and DL algorithms in wearables.

1.4 Problem Statement

The current studies have focused on improving the performance of speech communication devices such as ASR, VOIP, tele communication, tele conferencing etc. However, dealing with high level of noise in a noisy environment and providing noise-free communication is a trending research topic in this field. Several algorithms such as spectral subtraction, beamforming, adaptive noise reduction, spectral statistical filter etc., have been presented to improve the speech quality in MEMS microphones, but these algorithms suffer from low performance of recognition rate when signal to noise ratio is low (-15db, -10db, -5db, 0db) (Pauline, et al., 2021, Seon Man Kin et al., 2020). The MEMS microphone array consists of multiple microphones used to record audio/speech signals and provides the best speech recognition which is 71% Word Recognition Rate (WRR) at 10 dB SNR over a single microphone (Xu Yang et al., 2004; Alex stupakov et al., 2012).

Microphone arrays and speech enhancement are components built into Multi-Channel Speech Enhancement (MCSE) that processes multiple channels of audio signals in noisy environments such as outdoor (Alessandro et al., 2017; Pauline, et al., 2021). For example, a spectral statistics filter is applied to hearing aids for handling stationary noise environments (Gaussian noise) and unsteady noise environments (factories, babble and car noises) from -5 dB to 20 dB (Seon Man Kin et al., 2020). There is lack of improvement in performance rate of low SNR's resulting to a 2.162 PESQ (Perceptual Evaluation of Speech Quality) score with babble noise, where 2.203 is considered as low quality of signal with Gaussian noise, 2.133 considered as low quality of signal with factory noise and 3.677 PESQ score is considered as medium quality of signal with car noise on an average of -5db to 10db SNR level.

Existing MCSE systems can filter SNRs from 0 to 60 dB, providing 62.5% WRR at 0dB (considered as low SNR) and 83% at 60dB (considered as high SNR) of Gaussian noises (Alessandro et al., 2017). However, only White Gaussian stationary noise was tested between 0dB to 60 dB SNRs (Alessandro et al., 2017), and yet to be tested under nonstationary noisy environments such as airport noise, babble noise, car noise, exhibition noise, restaurant noise, factory noise, music noise, and helicopter noise. These non-stationary noises can get to very low dBs of SNR ranging -15dB, -10dB, -5dB.

Alessandro et al. (2017), used multichannel speech enhancement algorithm in MCSE which comprises of beamforming and adaptive noise reduction algorithms for filtering white gaussian noise. Speech communication devices have the most exposure to environmental noises used in outdoor environments, so there is a need to research MCSE under environmental noises. Regarding

advanced algorithms and feature extraction algorithms, there is a lack of research in implementing them in MCSE system.

Deep learning algorithms are considered as more advanced algorithms in the speech enhancement domain (Rownicka et al., 2017, Kinoshita, et al., 2020) which has been proven to offer acceptable performance in handling different levels of noises in speech enhancement that is based on a computing platform. Among the deep learning algorithms, VDCNN-conv reported the highest word recognition rate (WRR) at 90.45%, and the lowest WRR at 87.45% on an average for environmental noises (Pavani cherukuru et al., 2021). However, MCSE has never been investigated using Deep learning algorithms.

Recently most research make use of pre-processing approaches to obtain parameters from audio such as spectral parameters, temporal parameters etc., (Kanisha et al. 2018). Preprocessing algorithms help to improve the accuracy of recognition rate (Winursito et al. 2018) such as Mel-Frequency Cepstral Coefficients (MFCC), Principal Component Analysis (PCA), Linear Predictive Cepstral Coefficient (LPCC), Discrete Wavelet Transforms (DWT) and Wavelet Packet Transforms (WPT). According to past works (Takiguchi et al., 2007), without preprocessing of noisy speech signals, word recognition rate of noisy speech signals will amount to 63.9 %. Whereas, a combined usage of PCA and MFCC preprocessing algorithms will lead to increase in performance for noisy speech signals from 63.9% to 75.0%. Thence, there is a need to experiment the performance of preprocessing in the MCSE research area.

1.5 Research objectives

The major aim of this research is to improve the recognition rate at different SNR's of environmental noises in a Multi-Channel Speech Enhancement system. To achieve the research's aim, specific objectives have been identified as follows.

1. To analyse existing speech enhancement systems and multi-channel speech enhancement system in filtering different type of noises.
2. To experiment the performance of the existing multi-channel speech enhancement systems in handling environmental noises.
3. To develop a noise filtering framework using suitable algorithm(s) in multi-channel speech enhancement systems for filtering various Signal-to-Noise ratio (SNR) of environmental noises.
4. To evaluate the performance of the developed noise filtering multi-channel speech enhancement system in handling various Signal-to-Noise ratio (SNR) of environmental noises.

1.6 Research Questions

1. What are the components, algorithms used and the performance of the speech enhancement system and multi-channel speech enhancement system in filtering environmental noises? (objective 1)
2. What is the performance of existing multi-channel speech enhancement systems under noisy environments? (objective 2)
3. Which algorithm(s) is/are suitable to be applied on the proposed multi-channel speech enhancement framework in improving the performance at various Signal-to-Noise ratio (SNR) of environmental noises? (objective 3)

4. Can the proposed research improve the performance of the multi-channel speech enhancement in filtering environmental noises with acceptable results? (objective 4)

1.7. Research scope

- This research focuses on speech enhancement systems used in speech devices to incorporate the speech recognition system.
- Multi-Channel Speech Enhancement has been investigated in this research.
- The traditional algorithms are not suitable for low SNR noises, thus the research only focused on Gaussian noise. However, in real scenarios there are various noises with different levels of SNR. Thus, the existing schemes are not suitable for this type of scenarios. To overcome this, this research focused on developing a new scheme that can consider low SNR signals alongside focusing on different environmental noises such as Airport noise, Babble noise, Car noise, Exhibition noise, Restaurant noise, Street noise, Subway noise, and Train noise respectively. Moreover, this research considers improving the performance of visible speech enhancement.

1.8 Research Methodology

This research work is divided into four phases of study: Phase 1 analysed existing speech enhancement and Multi-Channel Speech Enhancement systems in handling different types of noises, Phase 2 experimented the performance of existing Multi-Channel Speech Enhancement systems under environmental noises at different levels of SNR, Phase 3 designed a noise filtering framework for MCSE in handling various SNR of environmental noises and Phase 4 evaluated

the performance of the developed MCSE system as depicted in Figure 1.3. Each phase is explained in detail as follow:

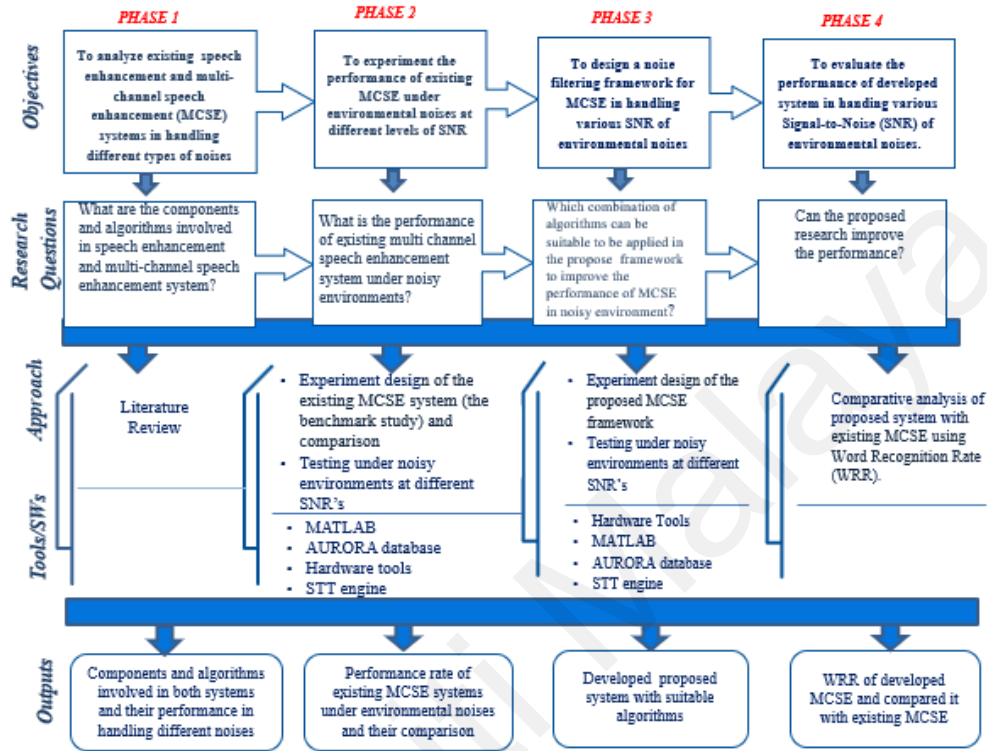


Figure 1.3: Research methodology

Figure 1.3 above, illustrates the research methodology of the thesis which is broken down as thus,

- Phase1: In phase 1, this research analyses the existing Multi-Channel Speech Enhancement (MCSE) systems in handling different type of noises from the literature review with the inclusion of journal article, conference papers, and magazines.
- Phase 2: In phase 2, the research conducts benchmark experiments on MCSE to identify the real problem existing in MCSE system. The activities include:

- (i) Experimental design of the existing speech enhancement systems (the benchmark study)
- (ii) Testing under noisy environments at different SNR's
- Phase 3: In phase3, the research develops a noise filtering framework using suitable algorithms tested in benchmark experiments for handling various SNR of environmental noises.
- Phase 4: Here, a comparison is made between the proposed MCSE system with the existing MCSE system based on Word Recognition Rate (WRR) accuracy.

1.9. Thesis organization

The complete thesis is organized into six chapters which are as follows:

Chapter 1 is the introductory chapter that describes several basics of speech processing, its applications, and noise sources in the original speech signal. This chapter also provide a brief discussion on speech enhancement and how signal quality for speech devices can be improved. Further, it describes the noise reduction algorithms such as single channel and multichannel speech enhancement alongside traditional speech enhancement algorithms. Based on this analysis, the study identified issues and challenges in this field and presented the significance of a deep learning scheme-based solution for Multi-Channel Speech Enhancement system.

Chapter 2 presented the literature review study where it described the existing algorithms of speech enhancement comprising, deep learning algorithms and multi-channel speech enhancement system. This research focused on single channel, multichannel speech enhancement, deep learning algorithms, preprocessing algorithms used for preprocessing the noisy speech signals and classification schemes.

Chapter 3 presented the experimental study of existing Multi-Channel Speech Enhancement system where a concise description was given on existing schemes such as deep learning algorithms, beamforming, adaptive noise cancellation, and voice activity detection. It also presented the outcome of implementation of VAD, ANR and beamforming algorithms for wearable speech enhancement. Finally, the chapter compared the MCSE systems under stationary and non-stationary noisy environments.

Chapter 4 presented the proposed noise filtering framework deep learning-based solutions to improve speech quality for speech communication devices, considering different types of noises and low level SNRs. Moreover, wavelet transform based pre-processing algorithm was also introduced to improve the overall performance of Multi-Channel Speech Enhancement system.

Chapter 5 presented the outcome of the proposed approach in terms of LLR (Log Likelihood Ratio), PESQ (Perceptual evaluation of speech quality), IS (Itakura-Saito), word recognition error (WRR) rate and SNR. The comparative analysis shows that the proposed approach achieved better performance when compared with the existing MCSE systems based WRR.

Chapter 6 presented the concluding remarks of this study as it presented novel solutions for Multi-Channel Speech Enhancement system and suggested future research work.

CHAPTER 2: LITERATURE REVIEW

2.1 Overview of this Chapter

This chapter gives a brief introduction to speech enhancement and noise in speech signals. It also includes detailed description and discussion of speech enhancement algorithms and various deep learning algorithms on speech enhancement. Furthermore, the chapter includes description of the main components that make up a speech recognition system where the focus is on speech enhancement for speech communication devices. A literature survey of the major algorithms developed by various researchers in each stage of recognition has been carried out as well. Thus, a clear understanding is provided on the developments that have taken place in each stage. Moreover, all available choices of speech enhancing algorithms are analysed based on their relative advantages and disadvantages. A comparative study of different dimension reduction algorithms for pre-processing and speech classification has been included in this chapter respectively. Consequently, the chapter presents detailed literature survey and review of the research papers and technical papers regarding speech recognition system in general, with concentrated emphasis in improving their performance. Finally, the performance evaluation methods for speech enhancement algorithms are explained at the end of the chapter.

2.2 Noise in Speech Signals

Speech, music, and noise from a variety of sources can all be combined (background noise, low-frequency noise, etc.) (Feng et al., 2022). It is assumed that the sound mixtures comprise at least one speech source that serves as the intended or target source for this thesis (i.e., speech augmentation). Different types of sources have distinct temporal and spectral characteristics, which are critical for various speech-enhancement algorithms (Wang et al., 2018; Tesch et al.,

2022; Kang et al.,2018). These are a few of the most important and common features of sound, music, and voice sources.

2.2.1 Speech sources

To express a certain message, speech is a sequence of sounds generated by the human vocal instrument system. Speech signals are distinct from other forms of signals because of their unique nature. The following are the primary features of speech:

2.2.1.1 Non-stationarity

Non-stationary stochastic processes can be used to model speech signals. The good news is that speech is thought to be quasi-stationary for short durations (of the order of 20 ms). In the analysis and production stages, a windowing procedure is required to allow for study of short-time signal sections as stationary processes. It is usual to use the short-time Fourier transform (STFT) (Takaki S et al., 2019) to accomplish this analysis.

2.2.1.2 Wideband signal

Speech signals have a range of about 7 kHz, but most of the data is contained within the band of frequencies up to 4 kHz in breadth. In comparison to the sampling frequency, this has a large bandwidth (usually 8 or 16 kHz). As a result, voice signals are wideband. Speech enhancement algorithms, especially those frequency media, must take this property into account when developing their algorithms (Pradhan et al., 2011; Abel et al., 2016).

2.2.1.3 Non-Gaussianity

Signals in speech are highly non-Gaussian, yet they are on the verge of having a PDF (Probability density function) that is near to being super-Gaussian. This attribute should be

considered by statistical algorithms and can be beneficial for the improvement of speech (Naik et al., 2012; Oliinyk, et al., 2020).

2.2.1.4 Human Speech production

The most widely accepted model of speech production assumes that speech is made up of two parts: an excitation signal and a filter associated to the vocal tract. Excitation signals are generated by a noise generator for unvoiced segments and by an impulse train generator for voiced segments that has the same period as the speech signal (the opposite of the fundamental frequency) (Yoneyama, et al., 2022). An AR (Autoregressive) model can be used to approximate the vocal tract filter. Figure 2.1 presents a visualisation of the speech production paradigm.

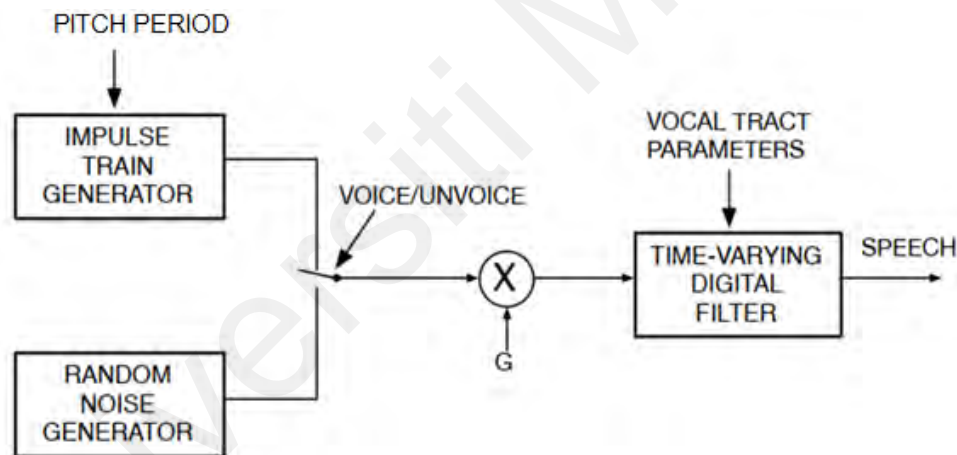


Figure 2.1: Speech production model (Edmund Lai et al., 2003)

2.2.1.5 Pitch

Pitch is a characteristic of every human voice since it is the subjective impression of the fundamental frequency (Willis et al., 2012). The average male voice is 140 Hz, while the average female voice is 200 Hz (Perry et al., 2001). The pitch of voiced sections varies throughout time, but it stays within a 40 Hz range (De Cheveigne, et al., 2002). To distinguish between distinct

speakers, pitch is an intriguing attribute to use in the process of voice source separation (Liu, et al., 2021).

2.2.2 Music sources

Musical instruments cause spectral discrepancies between speech and music spectra, which are often present in music transmissions. Speech usually has a well-defined range of perceptual properties that are well-established and predictable (Siedenburg et al., 2021). However, the spectral features of musical instruments substantially influence the spectral characteristics of the spectrum. In addition, the fundamental frequencies of music, ranges from 30 Hz to 4 kHz in frequency (Rob et al.,2000).

2.2.3 Noise sources

Airplanes, buses, cafes, cars, kindergartens, living rooms, the outdoors and classrooms, sports, traffic, trains, and train stations are just a few examples of the many noise sources that might obstruct clear speaking (Alexandre et al., 2018).

There are different kind of noises coming from our surroundings, which are referred to as environmental noise. Environmental noises can be categorised into two: i). Stationary noise: signal has constant frequency levels. For example, Additive White Gaussian Noise. ii). Non- Stationary noise: signals have different levels of frequencies in their frames. For example, train noise, crowded people, noise coming from car, babble noise when two or more people talk, airport surroundings, exhibition, restaurant, fan noise, vehicles noise, etc,. Non-stationary and stationary noise sources can be distinguished. Example of stationary noise can be the cabin noise of an airplane or the industrial noise of a manufacturing plant (Ryherd et al.,2008). The characteristics

of stationary noises is that the frequency and amplitude levels are constant in time domain modulation.

In addition to children's shouting in a classroom or babble noise, various non-homogeneous noises fall under the category of non-stationary noise. In terms of speech intelligibility, non-stationary noise has a greater impact than stationary noise, and it is harder to remove, because of its spectral characteristics (Baghel et al., 2020). Figure 2.2 depicts the spectrum differences between various types of noise, including a clear voice signal (a), babbling noise (b), train car noise (c), and white Gaussian noise (d).

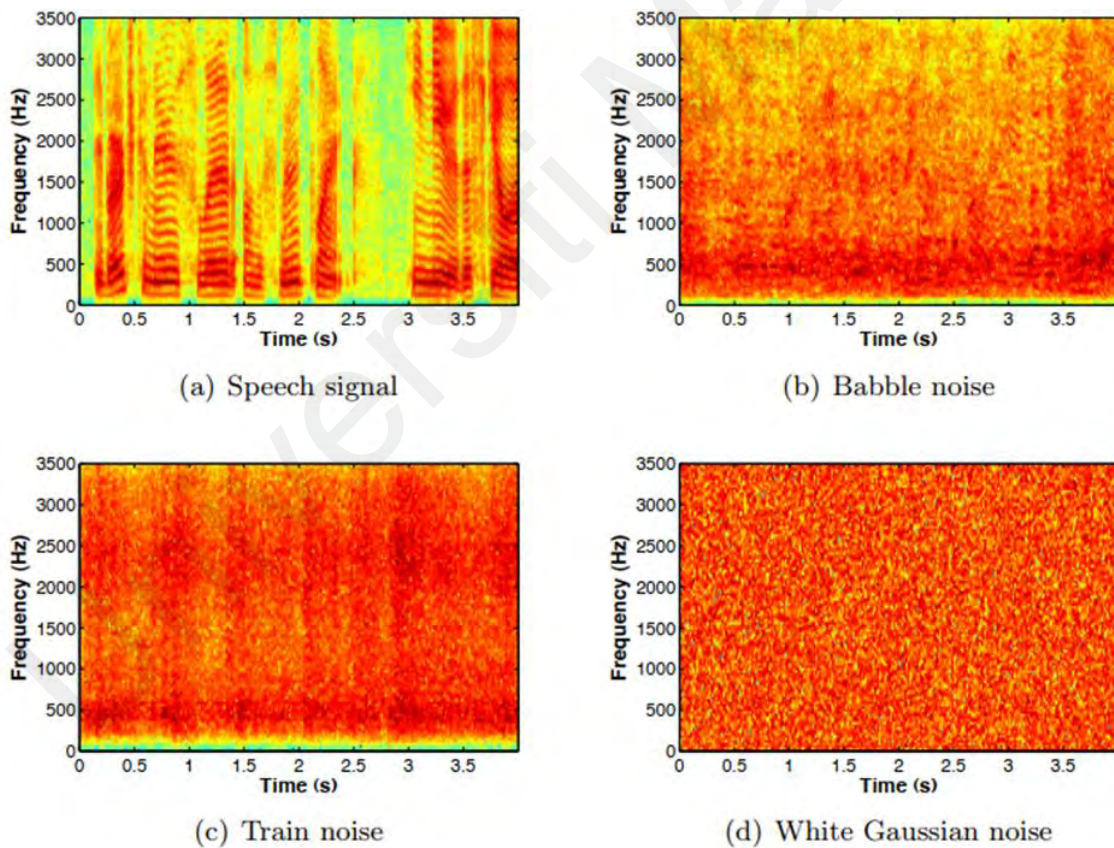


Figure 2.2: Spectrogram of speech signal (a) clean signal (b) babble noise (c) train noise and (d) Gaussian noise (Obtained from the Matlab simulation code)

The spectrograms of speech and noise signal as shown in Figure 2.2 are analysed at certain signal-to-noise ratios (SNR's). The signal-to-noise power ratio is expressed in decibels (Pinki et al., 2015). It is the most widely accepted and well-liked method for evaluating speech quality.

There are several factors of noise that affect the performance of speech processing and enhancement and these noises in speech signals are measured with SNR decibels. These SNRs varies from low level (etc....-10db, -5db, 0db 5db) to high level SNR's (10db, 15db, 20db.... etc.,). The following are the factors of noise affecting the performance (Nongpiur et al., 2013) such as:

- White noise, Coloured noise, Impulse noise, and Transient noise pulses are all examples of narrow-band noise (Emma Jokinen et al., 2014). Signal to noise ratio always varies from -10db to 20db.
- Background noise that is added to the speech signal, such as sound sources or engine sound when using a mobile phone. Here, the signal to noise ratio always varies from -10db to 20db.
- In a room with poor acoustics, an unintended echo can develop.
- Aural or acoustic feedback. An example is when the microphone of a two-way phone can catch a conversation between two people and relay both voices back to each other at the same time. Signal to noise ratio always varies from -10db to 20db.
- Due to the analogue signal's real value being rounded up, interference arises during sampling.
- Quality is lost.

2.3 Speech enhancement

High-quality speech signals and reliability against background noise, intervening sources, and reverberation effects are the goals of noise reduction algorithms (Hansler et al., 2008). The algorithms implemented in computer applications to enhance speech signals are referred as speech enhancement systems. Speech enhancement deals with noisy speech signals by reducing background noises while preventing alterations in speech features. This research used speech enhancement for speech signals processing applications like speech coder, automatic speech recognition, voice over internet protocol (VOIP), hearing aid, amongst others. Speech enhancement systems are frequently used as a pre-processor to enhance speech quality. Generally, algorithms used in speech enhancement systems consist of three types, namely filtering algorithms, spectral restoration algorithms, and speech model-based algorithms (Chen et al., 2008).

The filtering algorithms are used to filter the unwanted noisy signals that attenuate the noise features to produce a clean speech signal. Filtering algorithms contain time-domain filters, frequency-domain filters (Chen et al., 2008; Scalart et al., 2009; Hansler et al., 2006), and parametric filters (Chen et al., 2008). This research focused only on speech enhancement filtering algorithms.

2.3.1 Multichannel Speech Enhancement

Multi-channel speech enhancement can be implemented in both speech enhancement and Multi-Channel Speech Enhancement systems. If the system has several signal inputs, this work can employ adaptive noise cancellation devices, phase alignment to reject unwanted noise components, or even a combination of phase alignment and a noise-cancellation stage (Narine, et al., 2020) as shown in Figure 2.3. This is possible when the system has many signal inputs.

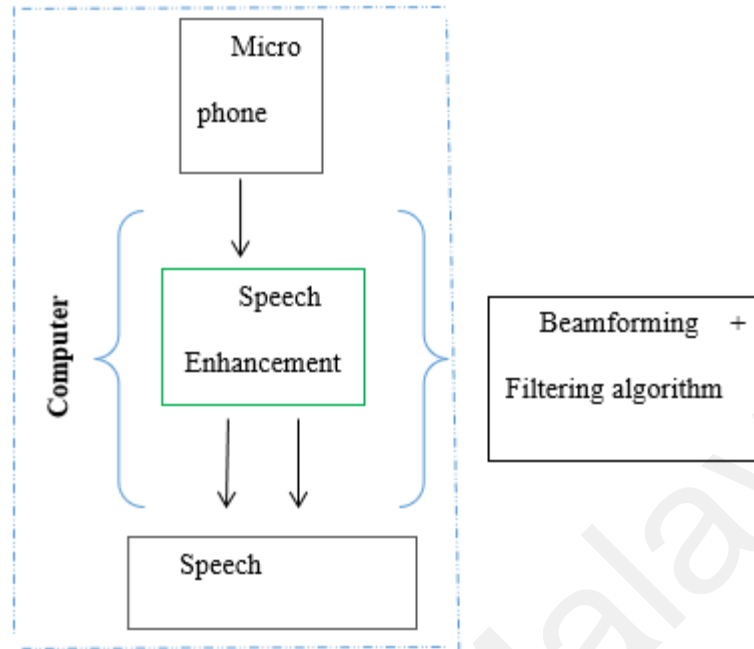


Figure 2.3: Generalised Architecture of Multi-channel Speech enhancement

This algorithm is more complicated than that of a single channel because it uses many channels. This uses noise reference in an adaptive noise cancellation device and makes use of the system's various signal inputs (Sugiyama, et al., 2019). By considering the spatial characteristics of the noise source and the signal as well as the constraints of the single-channel enhancement algorithm, this method can perform better for non-stationary noises (Yariv Ephraim 2018). These algorithms comprise two different methods: i). Multisensory beamforming and ii). Adaptive noise cancellation which are further explained thus:

- i) *Beamforming*: The most straightforward approach to multisensory beamforming using microphone arrays is drawn from radar and sonar applications and uses delay-and-sum beamforming. Because this work knows the direction in which the desired signal will arrive, it can make an assumption that the reflex ion's contribution will be minimal. So,

by aligning each sensor's phase function properly, this work may improve the intended signal while removing all the unwanted noise (Bactor et al., 2012).

- ii) *Adaptive Noise Cancellation*: When an auxiliary channel (referred to as the "reference path") is available, adaptive noise cancellation (ANC) can be used to reduce background noise and improve voice quality. This reference input will be filtered using an adaptive algorithm to remove the output of the filtering process, which contains noisy speech, from the main path of the analysis (Avalos et al., 2011).

2.3.1.1 *Multi-sensory Beamforming Algorithms used for speech enhancement*

Interfering and noise signals are encountered by spatially spreading signals. If the desired signal and noise are present in the same temporal frequency band, then the signals cannot be separated from interference signals using temporal filtering (Zhang et al., 2005). Generally, the desired and interfering signals arrive from different spatial locations. Therefore, spatial filtering can be applied to these signals to separate the desired signal from the interference using a microphone array known as beamformer (Elko et al., 2008).

A beamformer is a collection of sensors arranged in a specific way (Kumatani et al., 2019), such that the output of each sensor is processed before being joined together. Thus, like an FIR (finite impulse response) filter, a beamformer linearly mixes the spatially recorded waveforms of each sensor (Cheveign et al., 2010). Low-frequency arrays benefit from a significantly greater spatial aperture than a practicable single physical antenna, and they also benefit from the dimension reduction flexibility provided by discrete sampling when using an array of sensors (Dey et al., 2018). To completely suppress the interfering signal, the spatial filtering function is changed

in real-time applications but is impractical in the case of continuous aperture antennas (Vaughan et al., 2003).

Beamforming is mainly used in application areas such as RADAR, SONAR, Hands-free speech communication, image processing, biomedical and acoustic source localization. Microphone arrays have the capability of spatially sampling the sound pressure field (Rafaely et al., 2008). These microphone arrays are combined with spatio-temporal filtering known as Beamforming. When dealing with noise and other disturbing signals, one of the primary goals of beamforming is to predict the signal that will arrive from the intended direction. Since the frequency content of two signals that arrive from different directions overlaps, Beamforming can separate them. Microphone arrays seem to be the most promising noise reduction algorithm for solving the problem of reducing background noise and reverberation in a hands-free speech environment (Miyazaki et al., 2019).

Beamformers are classified into two categories: 1) data independent and 2) statistically optimum. All signal and disturbance signals must be able to respond to the beamformer's data-independent wavelet transformation. The weights in data-independent beamformer are designed such that the beamformer response approximates the intended result, regardless of the actual data or the statistics of that data (Wang et al., 2018).

The principle involved in designing these type of beamformers is similar to that of a conventional FIR filter Sum of the Delays Beamformer, which is also an example of a data-independent system (Galindo et al., 2020). Meddling and background signals interfere with spatially propagated signals. If the desired signal and noise are present in the same temporal frequency band, then the signals cannot be separated from interference signals using temporal

filtering (Xu et al., 2020). Generally, the desired and interfering signals arrive from different spatial locations. Therefore, spatial filtering can be applied to these signals to separate the desired signal from the interference using a microphone array known as a beamformer (Taseska et al., 2014).

A wavelet transform is a collection of sensors arranged in a specific way (Wang et al., 2018) such that each sensor's output is filtered before being joined together. Therefore, like a finite impulse response (FIR) filter, a wavelet transform linearly mixes the spatially recorded waveforms of each sensor (Saoud S et al., 2021). The advantages of using an array of sensors are: They can obtain a much larger spatial aperture than the practical single physical antenna in case of low frequencies and the other advantage is the spatial filtering adaptability offered by discrete sampling (Fischer et al., 2018).

These beamformers are used for the reduction of background noise in the acoustic environment as the user is at a distance from the microphone. It captures the background noise and interference due to the hands-free loudspeaker i.e., echo along with the desired signal (Reuven et al., 2007). In Wiener Beamformer, Correlation matrices for Signal and noise are calculated first, followed by calculation of the optimum weights. Elko's Beamformer is also implemented in which the normalized Least Mean Square (NLMS) algorithm is used for minimizing the error as well as to obtain a high correlation between the reference signal and desired signal in an acoustic environment (Gannot et al., 2008). Therefore, the performance of each beamformer is measured based on two parameters as Signal-to-Noise Ratio (SNR) and Speech Distortion (SD) as clearly explained in section 2.6.

(i) Elko's Beamformer

Noise level and reverberation can substantially damage the microphone receiving speech signals, which is a major issue in audio transmission for hands-free voice communication networks (Bertrand et al., 2011). Therefore, directional microphone arrays have the capability of solving both problems (Chen et al., 2015). One of the most popular microphones designed for asymmetrical microphone array is used in private communicators and videoconferencing (Meyer et al., 2008). In these type of arrays, sensors are near to the acoustic frequency in terms of separation. To find the direction, the microphone elements are placed in alternating sine fashion by Elko (2004). Therefore, it appears as a differential array due to the closed spacing between the microphone elements. This differential microphone is super-directional as the detector elements evenly summed output has a lower directivity than the measured one Zwysig (2009).

An orthotropic acoustic reverberation or sound field is assumed while designing a directed microphone. In real acoustic noise fields, it never reaches the theoretical assumption. Therefore, the solution to the above problem is to design an adaptive differential microphone system that results a directivity pattern which maximizes the signal-to-noise ratio (SNR) (Huang et al., 2019). Differential array sensors were not taken into attention as that of normal directional array sensors because the super-directional arrays are hard to realize. The Differential sensor array can be easily realized if the differential order of the sensor is limited to a first or second place (Benesty, et al., 2012). Therefore, the adaptive Differential microphone array is designed and implemented as low as possible while keeping in mind that a first-order microphone null is positioned in the back half plane (Moquin, 2004). As a result, even though this algorithm may not optimize the SNR in all acoustic settings, it greatly improves the SNR. To achieve SNR improvement and the purpose of microphone array, two unidirectional microphones were combined into back-to-back cardioid

microphones (Thomas, 2019). The weighted reduction of these two first-order array outputs can be achieved by combining two unidirectional microphones. An angular constraint can be applied to the null spot and some constraints can also be placed on the combination weighting (Thomas, 2019).

2.3.1.2 Maximum Signal-to-Noise Ratio (Max-SNR) Beamformer

The optimum beamformer which maximizes the output power ratio therefore is also known as Maximum array gain beamformer. The mean output signal of the beamformer is expressed as a function of filter weights (Araki, et al., 2007).

2.3.1.3 Delay and Sum Beamformer (DSB)

The basic idea of delay-and-sum beamformer is that when an electromagnetic signal arrives at the aperture of the antenna array, each element's output is added together with appropriate amounts of delays. Antenna array delays are determined by the physical distances between individual elements. The geometrical distance between the antenna elements and with each element are the parameters used to define the array features (Rakesh et al., 2017).

The interruption beamformer is a beamformer that does not rely on array data, such as the frequencies of the microphone sensors, for its response. As a result of the delay and sum as illustrated in Figure 2.4, beamforming delays are added after each microphone to allow for variances in voice signal arrival times to each microphone. Each microphone's time signals are put together. Various noise signals are blended in an unanticipated manner with signals that establish the intended speech signal. Because of this, the overall output's SNR is higher than the SNR of any single microphone signal. For the interruption beamformer, this work can conclude that it is more insensitive to sources in a specific direction (Wu, et al., 2019).

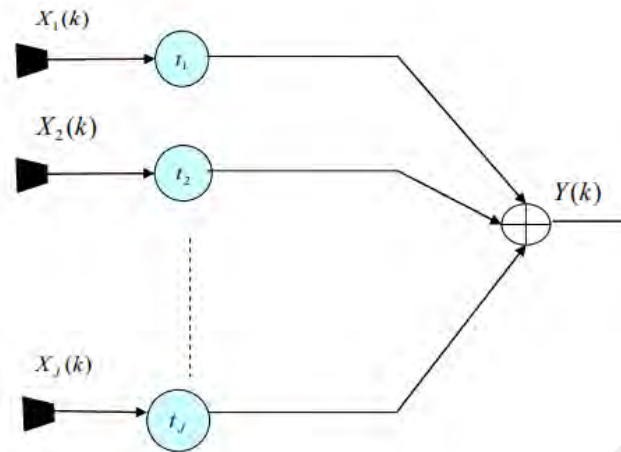


Figure 2.4: delay and sum beamformer with J microphones

The delay and sum beamformer, struggle from numerous issues. To improve beamformer SNR, it is necessary to use a high number of microphones, but this can only be done if incoming noise signals are totally statistically independent between microphones and the target speech signal. Also, nulls should not be positioned immediately in the path of incoming noise (Zeng, et al., 2013).

2.3.1.4 Adaptive Noise Reduction/ Cancellation Algorithms

Adaptive noise cancellation, phase alignment, or a combination of these algorithms can be employed in these systems, which take advantage of the many signal inputs available to the system (Kokkinakis and Loizou, 2010). Nonstationary noises can be best addressed by considering the signal's spatial features, as well as those of its noise source. These systems tend to be more complicated than others.

The following are the adaptive noise reduction algorithms based on the set of microphones:

(i) Derivative of Adaptive First-Order Array

There are two distinct ways to create first-order differential beamforming utilising two microphones. First, a single microphone signal is delayed alongside the difference between the resulting signal and secondly, the signal is then calculated. The so-called adaptive differential microphone array beamforming method is an alternative to this strategy that creates two cardioids and combines them to create the necessary spatial beam pattern.

- *Advantage:* Multiple noise sources with non-overlapping frequency content can be simultaneously eliminated by this processing since it is frequency-specific, even if the noise sources are in separate spatial locations (Buechner, et al.,2014).
- *Disadvantage:* Despite not requiring unique sensor technology, the differential approximation approach is known to have significant white noise gain, emphasising the significance of effective performance in the presence of noise. Additionally, the accuracy of a spatial derivative necessitates proximity of sensors in relation to the acoustic wavelength. This presumption falters at high frequencies, resulting in distorted beam patterns. Since distortion at the highest frequency of interest must be kept to acceptable or insignificant levels, the maximum distance must be set within an upper constraint when designing a differential vector sensor (Levin, et al., 2012).

(ii) NLMS based Adaptive First-Order Differential Microphone

The NLMS method's practical implementation is quite similar to the LMS algorithm because it is an extension of the ordinary LMS algorithm.

- *Advantage and disadvantage:* Due to its simplicity and stability, the algorithm is one of the most popular ones. The weak convergence is the only drawback (Malik, et al., 1991).

(iii) Least mean square (LMS) algorithm

Due to its simplicity and ease of use, the LMS or Stochastic Gradient algorithm is a frequently used adaptive algorithm. Therefore, this work created the back-to-back cardioid adaptive first-order differential array LMS method (Shah, 2020).

- *Advantage and disadvantage:* In contrast, this approach gives faster convergence but has more computing complexity.

(iv) Recursive least squares (RLS) algorithm

The adaptive filter algorithm recursively determines the coefficients that minimise a weighted linear least squares cost function pertaining to the input signals. This process is known as recursive least squares (RLS) (Stanciu, 2017). This strategy contrasts with other algorithms that try to lower the mean square error, including the least mean squares (LMS). The input signals are regarded as random for the LMS and other similar algorithms, but deterministic in the RLS derivation. In comparison to most of its rivals, the RLS displays incredibly quick convergence. This advantage is at the expense of considerable computational complexity.

- *Advantages:* The benefits of the recursive least squares (RLS) identification algorithm include straightforward computation and strong convergence characteristics (Malik et al., 1991).

Disadvantages: It is difficult to apply to data that has been censored. In comparison to maximum likelihood, it is often thought to have less desirable optimality qualities.

2.3.2 Deep Learning-Based Algorithms on Speech Enhancement System

Multi-channel speech enhancement is the process of using recordings from many microphones to eliminate reverberation, interference, and noise from a degraded speech signal. In traditional

methods, the signal from the target source is preserved while all other signals in the space are suppressed using linear spatial filters, such as those from a minimal variance distortion less-response (MVDR) optimization Furui (2018). Deep neural networks (DNNs), which are used for guided speech enhancement, have gained popularity in recent years (Karita, et al., 2019). Figure 2.5 depicts the architecture of speech enhancement system by using deep neural network (DNN) based algorithms.

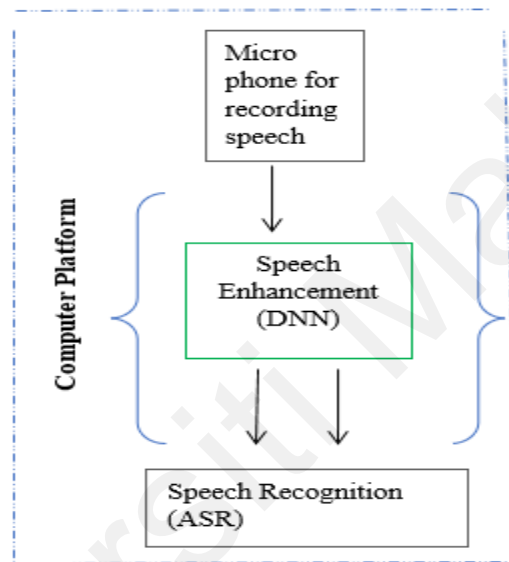


Fig: 2.5 Generalized Architecture of Deep Learning based Speech Enhancement

DNNs are typically combined with conventional spatial filters for multi-channel processing (Lind, et al., 2014; Meyer, (2018); Das, et al., 2021). Their purpose is to improve the spatial filter's estimates of speech and noise statistics. Training DNNs with spatial features such as inter-channel phase, duration, and level variations is another typical strategy (Karthik, et al., 2021; Andersen, et al., 2017). A DNN trained with spatial features is anticipated to use spatial cues for better target and interference discrimination.

Schroter et al. (2022) stated that deep learning-based speech enhancement and signal extraction have advanced due to complex-valued processing. The procedure is often based on the

application of a time-frequency (TF) mask to a noisy spectrogram, and complex masks (CM) are typically chosen over real-valued masks because of their capacity to change the phase.

One of the most popular architectures for Speech Enhancement is the deep neural network (DNN), also known as the feed-forward fully connected layer or multilayer perception (MLP) with several hidden layers (Zhao et al., 2018). Because every node in the layer shares a link with every node in the layer before it, the network is known as a completely connected network. DNN has relatively huge parameters as a result. The work by (Karjol et al., 2018) offered an enhancement algorithm using multiple DNN based system with n number of DNN, each of which contributes to the final enhanced speech, and utilising a gating network which gives the weights to combine the DNN outputs. Subjective and objective measures can be used to compare the performance of SE systems using the standard metrics. The model use n=4 with each layer being three layers deep. On the TIMIT corpus, an average SNR of -5 to 10 dB results in a seen noise perceptual evaluation of speech quality (PESQ) of 2.65 and an unseen noise PESQ of 2.19.

In the case of the worst signal-to-noise ratio, the processing is difficult and may cause signal distortions and a decrease in understandability, according to Dash et al. in 2020. This is done to improve the quality in speech and intelligibility, in reference to clear speech in terms of 0 (unintelligibility) and 1(excellent intelligibility). While overcoming the complexity of the current speech enhancement algorithms, a hybrid approach is put forth in this study. Using a modified deep neural network (DNN) and adaptive multi-band spectral subtraction (AdMBSS), the main goal of the research is to improve the intelligibility of the speech enhancement system that has been trained for a specific speech signal (Dash et al. 2020). AdMBSS is employed to increase the speech signal's understandability through the calculation of additional phase information, and to enhance the signal's quality, hybrid DNN and Nelder Mead optimization are applied. Although

DNN has been used successfully as a regression model for Speech Enhancement, the improved speech that it produces frequently degrades in low SNR situations (Gao et al., 2016).

To enhance the effectiveness of DNN-based speech in low SNR environments, some authors presented a progressive learning architecture with LSTM network (Gao et al., 2018; Santhanavijayan et al. 2021). Each of the target layers is built so that the transition speech with a higher SNR is learned at the last layer, followed by clean speech. Additionally, LSTM-RNN has been used to solve issues with reverberation (Weninger et al., 2013), multichannel loud speech (Li X et al., 2019) and extremely non-stationary additive noise (Wollmer et al., 2013). In (Wollmer et al., 2013), bottleneck features produced by the bidirectional LSTM network outperformed manually created features like MFCC (BN-BLSTM). When employing MFCC, the average word accuracy (WA) is 38.13%, whereas when using BN-BLSTM, it is 43.55%. The LSTM-RNN has significantly enhanced speech processing systems. However, it is well known that learning the RNN parameters is challenging and time-consuming.

Researchers in the field of speech technology has tailored focus on the convolutional neural network (CNN) and used the measurements to check the performance. Measurements include mean opinion score (MOS), signal distortion (SIG), and intrusiveness of background noise (BAK). A scale from 1 to 5 is used for SIG, BAK, and MOS, with a higher number being preferable. While word error rate (WER) or word accuracy (WA) is a common metric used to specifically assess the performance of ASR systems, other common objective measures include segmental signal-to-noise ratio (segSNR), distance measures, source-to-distortion ratio (SDR), perceptual evaluation of speech quality (PESQ), and short-time objective intelligibility (STOI). CNN produces the result as PESQ 3.24, CSIG 4.34, CBAK 4.10, COVL

3.81, and SSNR 16.85 respectively

(Wang et al., 2019; Park et al., 2017; Pandey et al., 2019; Germain et al., 2019; Fu et al., 2018; Rownicka et al., 2020; Rethage et al., 2018; Choi et al., 2019). Additionally, it has been said to be more effective than RNN (Park et al., 2017; van den Oord et al., 2016) and traditional feedforward neural networks (Fu et al., 2018). According to Park & Lee, CNN can perform better with a network that is 12 times smaller than RNN (Park et al., 2017). CNN is effective in distinguishing between the speech and noise components of noisy signals because it can handle the local temporal spectral features of speech. Both in the spectrum and waveform domains, CNN has demonstrated its efficacy for improving speech. Table 2.1 presents existing studies on deep learning algorithms and dataset used in their research experiments, metrics used for evaluation, results, advantages, and disadvantages.

Table 2.1: Existing studies on Deep learning algorithms alongside their metrics, data base used, results, advantages and disadvantages.

Deep Learning Method	References	Dataset	Evaluation Metrics	Results	Advantages/Disadvantages
DNN (Deep Neural Network)	Zhao et al., 2018	NOISEX and IEEE corpus	SDR, PESQ, and STOI	Averaged results with mismatched SNR (-3 to 3 dB) PESQ is 1.99, SDR is 11.35, and STOI is 90.61%.	Advantages being familiar with the model's architecture since Networks are typically simple. Disadvantages DNN has relatively big parameters since every node in each layer is connected to every node in the layer before it.
	Bhagachi et al., 2018	CHiME-2	WER	Error rate of 14.7%.	
	Karjol et al., 2018	TIMIT + noises from Aurora dataset	STOI, SegSNR, and PESQ	For seen noise, the average best PESQ is 2.65, whereas for unseen noise, it is 2.19.	

	T. Gao et al., 2016	WSJ + environmental and musical noises	PESQ, STOI, and SSNR	PESQ was assessed on unseen noise and was 1.93 for single-SNR training and 1.82 for multi-SNR training.	
Deep autoencoder based on MFCC (DAE-MFCC)	X Feng et al., 2014	CHiME-2	WER	Error rate of 34%.	Advantages Dimensional reduction is done using DAE, and the bottleneck layer's features might be helpful.
	X Lu et al., 2013	Japanese corpus + environmental noises	PESQ	Average PESQ for factory noise is 3.13, whereas it is 4.08 for car noise.	
Recurrent neural network-Long short-term memory (RNN-LSTM)	T. Gao et al., 2018	In factories, the average PESQ is 3.13, and in cars, it is 4.08.	SDR, STOI	STOI: 0.86 and SDR: 9.46 on average.	Advantages -Best for handling data that is sequence-based, like speech signals. -Contextual data can be handled by RNN-LSTM.
	F. Weninger et al., 2013	CHiME-2	WA, WER	Average accuracy is 85%.	
	M. Wollmer et al., 2013	Buckeye (spontaneous speech) + CHiME noises	WA	Average WA using BN-BLSTM: 43.55%.	
	A L maas et al., 2012	AURORA-2	MSE and WER	The average error rate (SNR 0-20 dB) is 10.28% for seen noise and 12.90% for unseen noise.	
	P Wang et al., 2019	CHiME-2 + environmental Noises	WER	Magnitude features provide the best average error rate of 7.8% (accuracy of 92.2%).	
					Disadvantages Learning temporal information is a drawback of DNN-based DAE information.
					Disadvantages It is well known that learning the RNN parameters is challenging and time-consuming.

S R Park et al., 2017	TIMIT + environmental noises	PESQ, STOI, SDR	CNN outperformed DNN and RNN in terms of accuracy, with PESQ 2.34, STOI 0.83, and SDR 8.62.
P. Plantinga et al., 2019	CHiME-2	Word Error Rate (WER)	Using ResNet and mimic loss, a word error rate of 9.3% is achieved.
J. Rownicka et al., 2020	AMI and Aurora-4	Word Error Rate (WER)	8.31% WER on Aurora-4
A. Pandey et al., 2019	NOISEX + TIMIT + SSN	STOI, PESQ, and SI-SDR	Results indicate that Autoencoder CNN performed better than SEGAN.
F. G. Germain et al., 2019	Voice Bank + DEMAND	SNR, SIG, BAK, OVL	SNR:19.00, SIG: 3.86, BAK: 3.33, OVL: 3.22.
S. W. Fu et al., 2018	TIMIT + environmental noises	PESQ, STOI	Fully utilising ConvNet yields the best STOI, while DNN achieves the best PESQ.
D. Rethage et al., 2018	Voice Bank + DEMAND	DEMAND SIG, BAK, OVL, MOS	3.60 MOS is achieved. When compared to the Wiener filter, overall outcomes are superior.
C Donahue et al., 2018	WSJ + environmental and Music noise	Word Error Rate (WER)	17.6% word error rate.
D. Baby et al., 2019	Voice Bank + DEMAND	STOI, PESQ, SegSNR	PESQ: 2.62, SegSNR: 17.68, STOI: 0.942

	S. Pascual et al., 2017	Voice Bank + DEMAND	PESQ, CSIG, CBAK, COVL, SSNR	PESQ is 2.16, CSIG is 3.48, CBAK is 2.94, COVL is 2.80, and SSNR is 7.73 are the values.	
CNN (Convolution neural network)	K. Kinoshita, T et al., 2020	CHiME-4, Aurora-4	WER, SDR	Chime-4: SDR: 14.24, Aurora-4: 6.3%, WER: 8.3% (real data), 10.8% (simulated).	Advantages -CNN has the capacity to detect patterns in neighbouring speech structures. -Compared to RNN and standard DNN, CNN is more effective. Disadvantages inability to maintain invariance when the input data changes
	Z. Xu et al., 2020	Grid corpus + CHiME-3 noises	PESQ, STOI	For seen noises, PESQ is 2.60 and STOI is 0.70, while for unseen noises only, 2.63 and 0.74.	
	H. S.Choi et al., 2019	Voice Bank + DEMAND	PESQ, CSIG, CBAK, COVL, SSNR	PESQ 3.24, CSIG 4.34, CBAK 4.10, COVL 3.81, and SSNR 16.85 are the values.	
GAN (generative adversarial network)	M.H.Soni et al., 2018	Voice Bank + DEMAND	PESQ, CSIG, CBAK, MOS, STOI	PESQ 2.53, SIG 3.80, BAK 3.12, MOS 3.14, and STOI 0.93T are the values.	Advantages: If GAN is correctly trained, its combined networks can be very strong. Disadvantages: The adversarial training is typically challenging and unstable.
	A. Pandey et al., 2018	TIMIT + NOISEX + SSN	SSN, PESQ, STOI	GAN gives consistently better STOI score, but not much of an improvement in PESQ.	

2.3.3 Summary of Speech enhancement Algorithms

- Recent advance of deep learning technologies has provided great support for the progress in speech enhancement research field. Unlike conventional speech enhancement approaches that depend on statistical model, deep learning approaches build on a data-driven paradigm.
- Conventional speech enhancement such as spectral subtraction, Wiener filtering, and minimum mean square error, have been outperformed by deep learning methods. The development of deep learning is one of the most significant technologies today.
- In deep learning methods, RNN performs (8.31% WER on Aurora-4) better than DNN.
- CNN is the advanced neural network of DNN, it gives better performance in terms of WER, SDR, PESQ, STOI, SIG, CBAK, COVL AND SSNR (explained in section 2.6). Subjective and objective measures can be used to compare the performance of SE systems using the standard metrics. Common subjective measurements include mean opinion score (MOS), signal distortion (SIG), and intrusiveness of background noise (BAK). A scale from 1 to 5 is used for SIG, BAK, and MOS, with a higher number being preferable. While word error rate (WER) or word accuracy (WA) is a common metric used to specifically assess the performance of ASR systems, other common objective measures include segmental signal-to-noise ratio (segSNR), distance measures, source-to-distortion ratio (SDR), perceptual evaluation of speech quality (PESQ), and short-time objective intelligibility (STOI). The higher the value of PESQ, which ranges from -0.5 to 4.5, the better the speech quality.
- CNN has the capacity to detect patterns in neighbouring speech structures.
- Compared to RNN (8.31% WER on Aurora-4) and standard DNN (14.7% WER on CHiME-2, CNN (6.3% WER on CHiME-4 and AURORA database) is more effective.

2.4 Multi-Channel Speech Enhancement

Single microphone systems only make use of the received signal's spectral and temporal variety. Spatial diversity is also induced by reverberation. The use of numerous microphones is necessary to fully take advantage of this diversity. To improve speech captured by many microphones, beam forming-based spatial spectrum estimation algorithms have been used in the literature. Multi-channel systems employ several signal inputs to the system as well as noise reference in an adaptive noise cancellation device in the context of noise cancellation.

The multi-channel system also employs phase alignment to filter out unwanted noise components. Thus, non-stationary noises can be better addressed by utilizing the spatial characteristics of the signal and the noise source. As a result, the drawbacks of one channel systems are overcome. Due to an increase in hardware requirements, multi-channel systems have complex structures and are expensive. However, when compared to single channel systems, multi-channel systems provide greater outcomes for speech enhancement.

Single-channel recognition results in poor speech recognition when the speaker is distant from the recording microphone device Lind, et al., 2014 (typically of something more than 0.2m). Also, single-channel approaches are affected by low SNR and high reverberation conditions. This resulted into the popular use of microphone array alongside the added advantage of using its strategic microphone placement, in obtaining spatial information. Firstly, a microphone array can locate and track a speaker since different positions of speakers produce different instances of this signal being received at the microphones. Secondly, simultaneous source signals overlapping in frequency domain but coming from different directions can be separated using such an array. Microphone arrays can steer its response in different directions, allowing it to extract the signal from a particular direction attenuating other signals from other directions, which is called

beamforming. Numerous narrowband array processing algorithms have been adapted or generalised (in a very straightforward manner) for use in microphone array processing. This has the benefit that most algorithms created for antenna arrays decades ago can be expanded with little effort (Meyer, (2018)). In antennas, array processing is used for directional reception as well as transmission of narrowband signals. So much of the theory behind the construction of spatial filters were derived from these narrowband processing algorithms. Since speech is a wideband signal, most of the array processing algorithm works by considering each frequency bin as a narrow band signal and applying the narrowband algorithms to each bin.

There are different components of Multi-Channel Speech Enhancement based on multi-channel which helps to filter the noise by capturing signals from MEMS microphones as shown in Figure 2.7:

- Beam forming: Grouping the different signals based on DMA theory (microphone).
- Active Noise Reduction: The ANR is an adaptive filter aimed to delete environmental noise using an adaptive LMS filter.
- Voice Activity Detection: The purpose of this VAD algorithm is to discriminate the presence of the user's voice in audio stream.

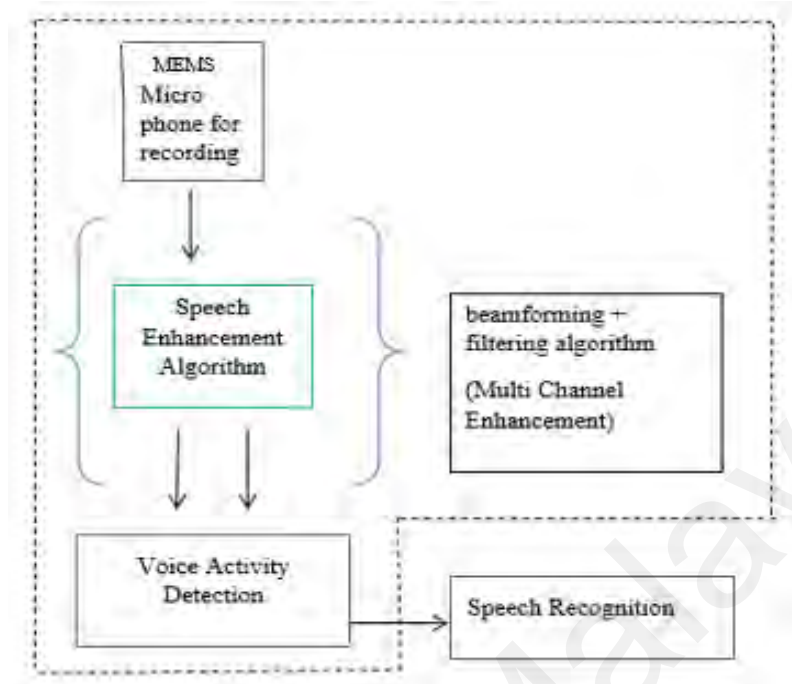


Figure 2.7 Generalised Architecture of Multi-Channel Speech Enhancement (Alessandro et.al., 2016)

Speech software applications (such as ASR) must reduce noise if they are to increase performance and resilience when wearable devices are utilized in noisy environments. For example, the signal-to-noise ratio can be greatly improved by having everyone wear a portable microphone near their mouth instead of utilizing a remote microphone.

2.4.1 The performance of the Multi-Channel Speech Enhancement systems in the existing studies

To improve the signal-to-noise ratio (SNR) of the user's voice, Alessandro Palla et al. (2015) suggested a speech enhancement system based on a MEMS (Micro-Electro-Mechanical Systems are a more attractive choice in the use of noise cancellation) microphone array and a digital signal processor. The array uses a Differential Microphone Array (DMA) and an Adaptive Noise Reduction approach to take advantage of the audio delay between microphones. When sound and

voice are coming from different directions, the system can achieve an increase in SNR of around 16.5 dB. When a user talks, a voice activity detection (VAD) block detects it and sends the information to an ASR system that is cloud-based. The embedded system can be incorporated into a speech communication device because of the modest array size.

The effectiveness of Micro Electro-Mechanical System (MEMS) microphones, a recently developed technology with very small sensors, for multichannel voice enhancement is examined by Skordilis et al. (2015). Real-world speech data gathered with a MEMS microphone array is used in experiments. A new corpus of voice recordings made in diffuse and localised noise fields utilising a MEMS microphone array set up in linear and hexagonal array geometries is used to first investigate the efficacy of the array shape for noise reduction. This paper shows that the hexagonal geometry performs better. The ATHENA database, which contains speech captured in accurate smart home noise settings using hexagonal-type arrays of both microphone types, is then used to compare MEMS microphones to Electret Condenser Microphones (ECMs).

An adaptive speech enhancement approach was introduced by Rao et al. (2016) that operates in real-time on Android devices to enhance the intelligibility and understandability of speech for hearing aid users. This research took advantage of a two-microphone speech enhancement approach as the first stage of processing, after which the second processing stage was then used, which is a modified single microphone SE method. To increase the quality and understandability of improved speech, a tuning factor is added to the calculation of a-priori signal-to-noise ratio (SNR). Users can adjust the enhancement for various types of background noise using the suggested SE method's graphical user interface (GUI), which is implemented on an Android platform and runs in real-time.

According to Sun et al. (2020), a supervised speech enhancement algorithm powered by a regression-based RNN (Recurrent Neural Network) structure is suggested to increase the performance of hearing aids that are mostly used with smartphones. This method overcomes the computing resource limitation of the traditional ear-worn hearing aid devices by utilising the powerful processor and smart-phone based architecture of the mobile platform. To considerably improve speech intelligibility under low SNR settings, a general challenge for improving speech intelligibility is investigated. The structure of the post-filters is also used to manage the specifics of acoustic signals. To meet the need for real-time processing with the constrained computational power of smartphones, the overall computation complexity is purposefully kept low. This approach is used on a smartphone to confirm its viability and demonstrate its improved performance. However, the PESQ and STOI results are average which is 68.5% of STOI and 1.51 of PESQ at -5db SNR, yet there is a need for improving the performance.

Gyuseok Park et al. (2020) introduced Speech Enhancement for Hearing Aids with Deep Learning on Environmental Noises. In this study, the speech enhancement for hearing aids was examined in real, noisy surroundings that were self-recorded. Convolutional networks were used to categorise environmental noises to improve voice quality, and DNNs were then used to apply noise reduction based on the classified noise. The PESQ, STOI, OQCM (overall quality composite measure), and LLR scores were used to objectively assess the improvement in speech quality in the ten locations where environmental noise was most closely associated to the setting in which hearing aids are used. When the classification findings were not used, the PESQ score rose by, correspondingly, 2.17% at 0 dB SNR, 3.50% at 5 dB SNR, 3.69% at 10 dB SNR, and 2.62% at 15 dB SNR testing. In the tests with 0 dB SNR, 5 dB SNR, 10 dB SNR, and 15 dB SNR, the STOI score increased by 3.23%, 2.71%, 1.89%, and 1.30%, respectively. The OQCM increased by 0.203

in the 0 dB SNR test, 0.243 in the 5 dB SNR test, 0.225 in the 10 dB SNR test, and 0.161 in the 15 dB SNR test, respectively. It was estimated using a mixture of the existing objective assessment methods. Table 2.2 presents the existing studies of Multi-Channel Speech Enhancement with different factors.

Table 2.2: Multi-Channel Speech enhancement algorithms using MEMS microphones

References	Components	Factors	Noises	Limitations
Alessandro Palla et al.,2015	<ul style="list-style-type: none"> • Microphone array • Beamforming • Adaptive Noise Reduction 	<ul style="list-style-type: none"> • Recording • Placement of Microphones • Signal directions • Filter noisy signals using filtering algorithm. • Segmentation • Data Reduction 	White Gaussian Noise	<ul style="list-style-type: none"> • WRR is poor at low SNR. • Tested with only Gaussian noise
Sumit Basu et al., 2000	Microphone array	<ul style="list-style-type: none"> • Recording • Placement of Microphones 	Acoustic noise	<ul style="list-style-type: none"> • WRR is low. • Close talking microphone is not at all satisfactory to improve recognition rate in noisy conditions
Yong Xu et al., 2004	Microphone array	<ul style="list-style-type: none"> • Recording • Placement of Microphones 	Real time noise	<ul style="list-style-type: none"> • Recognition rate is low

2.4.2 Summary of Multi-Channel Speech Enhancement

- The existing studies have focused on improving the performance of these speech communication devices.
- However, dealing with high level of noise in noisy environment and providing the noise-free communication is a hot research topic in this field.
- Several algorithms have been presented to improve the speech quality in MEMS microphones, but these algorithms have issue of low performance.
- It was tested only with white Gaussian noise but never tested with environmental noise. For instance, Alessandro et al. (2017) did not consider all types of noises in a real-time environment.
- MCSE in using MEMS microphones never investigated Deep learning algorithms and pre-processing algorithms.

Due to lack of pre-processing and deep learning algorithms, the performance is low.

2.5 Speech Enhancement for Automatic Speech Recognition

2.5.1 Pre-processing

The preprocessing algorithms are used to eliminate redundant data from an input speech signal. The two primary kinds of preprocessing algorithms depend on the type of parameters that needs to be obtained. Spectral parameter analysis methods use the spectral representation of the speech signal and temporal parameter analysis method, which make use of a signal's original format. Mel-Frequency Cepstral Coefficients (MFCC), Principal Component Analysis (PCA) and

Linear Predictive Cepstral Coefficient (LPCC) are all categorised under spectral parameter analysis. Whereas Discrete Wavelet Transforms (DWT) and Wavelet Packet Transforms (WPT) are classified under temporal parameter analysis (Maria Labied et al., 2021).

Many studies used the reprocessing algorithm based on MFCC parameters. Since the mid-1980s, MFCCs have been the most popular algorithm in the ASR community. The MFCC has been employed in the majority of Moroccan Darija speech recognition works (Mouaz et al., 2019; Ezzine et al., 2020). The primary function of PCA in the pre-processing stage is to determine a linear combination that may be utilised to represent the original speech signal. It is the most popular algorithm for boosting speech recognition systems' robustness in noisy environments. According to the research discussed in (Veis et al., 2011), the PCA analysis is necessary when the voice signal has been distorted by noise. Another study supports the finding that using PCA further reduced error rates (Lee J Y et al., 2011). According to the study (Takiguchi et al., 2007), using PCA and MFCC together increased identification rates for noisy voice signals from 63.9% to 75.0%. Several research measured LPCC and MFCC performance. The outcomes from (Venkateswarlu et al. 2011) demonstrate that MFCC and LPCC produce the same outcomes. Another study evaluating the two algorithms (Li et al., 2007) found that LPCC was 5.5% faster and 10% more efficient than MFCC.

For speech recognition applications, the temporal information in speech signals is just as significant as the frequency information (Sak et al., 2015). Due to the non-stationary nature of speech signals, DWT uses the mother wavelet to re-scale, shift, and analyse temporal information. The input voice signal is evaluated in this way at different frequencies and resolutions. The DWT provides an ideal model for the human auditory system because a speech signal is examined at

diminishing frequency resolution at rising frequencies, and it has been utilised in several studies at the preprocessing stage. According to Maria Labied et al., 2021, DWT preprocessing algorithm is rated as 5 based on weighted scoring method (WSM) in terms of the robustness to noises.

An addition to the basic wavelet decomposition that offers more signal processing features is the wavelet packet transform. It represents high-frequency information better than the wavelet transform. Wavelet packet transforms (WPT) split details and approximations, which is their significant differentiation from wavelet transforms. In comparison to DWTs, WPTs have more decomposed approximation and detail coefficients. The study of Iosif et al., 2007 compared DWT's performance to WPT's performance for the task of ASR, and the results revealed that DWT-based approaches outperformed WPT's.

Table 2.3 Comparison of preprocessing algorithms based on various measuring factors

References	Feature Extraction	Comparison with most important measuring factors					Advantages	Disadvantages
		Robust to Noise	Memory storage	Dimensionality reduction	Computational complexity	Computational speed		
Katti et al., 2011, Winursito et al., 2018 and Maria Labied et al., 2021	MFCC	Very poor at noisy speech signals	Inefficiency and a lack of flexibility in the amount of storage space required to store important details from a speech while performing spectral analysis on it.	Inefficiency and a lack of flexibility in achieving a high degree of accuracy while reducing the dimensionality of the extracted features.	Approved effectiveness of time and speed-related computational costs of a preprocessing algorithm	Good choice but there are some limitations in computational costs of preprocessing in terms of speed and time.	<ul style="list-style-type: none"> • High accurate recognition when noise free. • High discrimination and low correlation coefficients 	Recognition accuracy is very poor under noisy conditions.
Veisi et al., 2011; Katti et al., 2011; Grama et al., 2017	PCA	Approved effectiveness at noisy speech signals	Significant results are attained in the amount of storage space required to store important details from	Approved effectiveness in achieving a high degree of accuracy while reducing the dimensionality of the obtained parameters.	Inefficiency and a lack of flexibility of time and speed-related computational costs of a preprocessing algorithm	Significant results are attained in computational costs of a preprocessing algorithm in terms of speed and time	<ul style="list-style-type: none"> • Resistance to noises 	<ul style="list-style-type: none"> • High bandwidth computation is expensive.

			a speech while performing spectral analysis on it.					
Venkateswarlu et al., 2011 ; Katti et al., 2011	LPC	Very poor at noisy speech signals	Good choice but there are some limitations in the amount of storage space required to store important details from a speech while performing spectral analysis on it.	Significant results are attained in achieving a high degree of accuracy while reducing the dimensionality of the extracted features.	Good choice but there are some limitations of time and speed-related computational costs of a preprocessing algorithm.	Approved effectiveness in computational costs of preprocessing in terms of speed and time	<ul style="list-style-type: none"> • Computational speed • Robust for extracting features from speech samples with a low bit rate 	coefficients of highly related features
Ping et al., 2009; Katti et al., 2011; Maria Labied et al., 2021	DWT	Approved effectiveness at noisy speech signals	Significant results are attained in the amount of storage space required to store important details from a speech	Approved effectiveness in achieving a high degree of accuracy while reducing the dimensionality of the obtained parameters.	Inefficiency and a lack of flexibility of time and speed-related computational costs of a preprocessing algorithm.	Significant results are attained in computational costs of preprocessing in terms of speed and time.	<ul style="list-style-type: none"> • Speech signal denoising • Speech signal compression without significantly degrading its quality. 	Inflexible

			while performing spectral analysis on it.					
--	--	--	-------------------------------------------	--	--	--	--	--

Among all preprocessing algorithms such as Mel-Frequency Cepstral Coefficients (MFCCs), Discrete Wavelet Transforms (DWTs), Linear Predictive Coding (LPC), DWT performance is very effective in terms of denoising the speech signal and compressing speech signal without any significant loss in speech quality (Ping et al., 2019 , Katti et al., 2011 and Maria Labied et al., 2021).

Universiti Malaysia

2.5.2 Speech Classification

The process of improving speech signals involves reducing or eliminating additive noise. Numerous applications, including powerful speech recognition, teleconferencing, and hearing aids, adopt it as a pre-processor. Traditional algorithms for improving speech include nonnegative matrix factorization, Wiener filtering, and spectral subtraction algorithms. There are several uses of speech enhancement, including automatic speech recognition (ASR), headphones, VoIP (Voice over IP) communication, mobile communication systems, and hearing aids.

The problem of speech enhancement has long captured the attention of signal processing researchers, but it has never been fully resolved. Numerous approaches have been put forward to address this difficult task, ranging from the traditional approaches first put forth in the 1970s, which are based on statistical hypotheses about the noise present in the speech signal, to the more advanced approaches researchers have reached today, based on deep learning algorithms.

The traditional methods, which have been around for a while, are based on statistically analysing the relationship between speech and noise. Although some of these strategies were claimed to be successful in improving noisy speech, it has been demonstrated that these algorithms work best in surroundings with a high Signal to Noise Ratio (SNR) or in situations where stationary noise conditions exist. Additionally, it was asserted that these methods don't assist in making speech more understandable. To learn the mapping function that provides the best prediction of the clean speech without making any statistical assumptions, a Deep Neural Network (DNN) is trained using pairs of clean and noisy speech signals in a deep learning-based supervised speech enhancement.

According to Ma et al. (2021), feature learning and sample learning are both included in machine learning. Classification accuracy can be improved using deep learning (deep feature learning) by generating high-level and high-quality features through deep feature transformation. Some samples are of poor quality for classification due to data collection issues. Sample learning is therefore required. In contrast to sample selection, deep sample learning generates high-level and high-quality data through deep sample modification. Deep sample learning research, on the other hand, has never been made public. This challenge is addressed in past research by designing a deep dual-side learning ensemble model. The deep dual-side learning of PD (Partial Discharges) voice data is achieved in the existing model by designing and combining a deep sample learning algorithm with a deep network (deep feature learning).

Zhao et al. (2016) mentioned that speech identification from a distance is difficult because of the reverberation that occurs when speakers and microphones are separated by a significant distance. An ensemble of deep neural networks (DNNs) and a joint ensemble of DNNs were introduced to deal with the vast spectrum of reverberations that occur in real-world settings. As a preliminary step, differing reverberation times are taken into consideration while designing numerous DNNs. As an additional step, the ensemble of DNN acoustic models includes a feature mapping component that is intended to be used as a front-end for dereverberation. For testing, an ensemble of DNNs is combined using convolutional neural network estimates of prediction probabilities weighted averages (CNN). In other words, the DNN posterior probability outputs are blended using CNN-based weights as a weighted average of the DNN posterior probabilities.

2.6 Theoretical Background and Implementation Requirements

2.6.1 Discrete Wavelet Transform (DWT)

Wavelets can be created by rescaling and iterating through a series of filters. Up-sampling and down-sampling (subsampling) processes determine the signal's resolution (detail information), whereas filtering operations determine its scale (resolution).

Figure 2.8 illustrates how the discrete-time-domain signal is processed to compute the DWT using lowpass and highpass filtering, respectively. The Mallat algorithm or Mallat-tree breakdown is the technical term for this process. It is important because it links the time series multiresolution to discrete time filters. The sequence $x[n]$, where n is an integer, is used to represent the signal in the image. It is possible to represent a low pass filter as G_0 while low pass filter is designated by L_0 , whereas the high pass filter is designated by H_0 . Fine details are generated at every level by the high pass filter, $d[n]$, while the scaling function's low pass filter provides coarse approximations, $a[n]$.

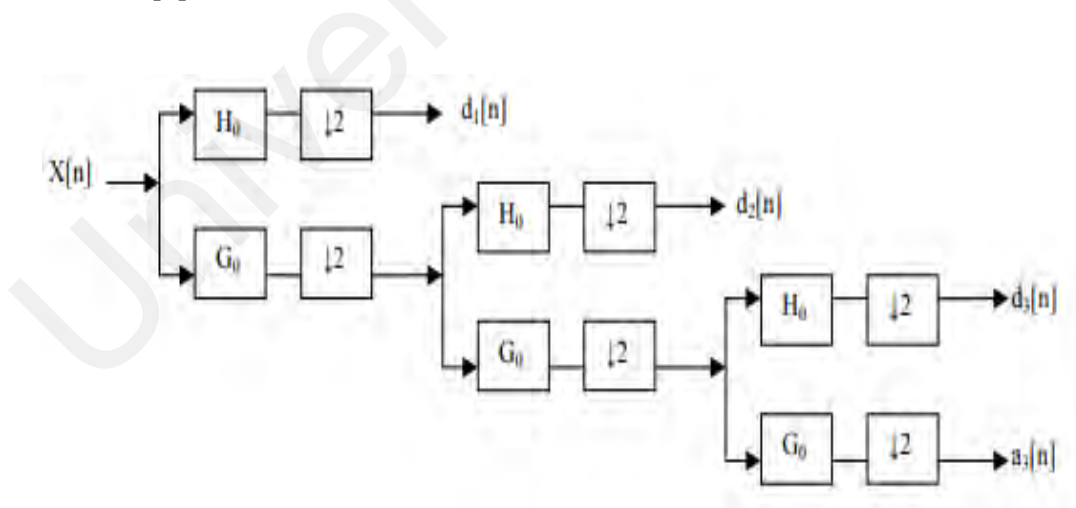


Figure.2.8: Wavelet transform (Hazrat et al., 2014)

The frequency resolution in higher levels is improved by decomposing the breakdown just into the estimate component at each level using the triadic filter bank. Regular wavelet analyses may yield lower DWT decompositions. Certain applications may experience difficulties while using DWT because of the importance of the information in the higher frequency components. It is possible that the decomposition filter's frequency resolution is not precise enough to get the information this research needs from the signal's deconstructed component. A wavelet packet transformation can be used to further breakdown a signal and attain the desired frequency resolution. Like the DWT, the wavelet packet analysis also decomposes a wavelet detail component into its own estimation and detail components, in addition to the wavelet estimation component at each level.

The bandwidth of a filter narrows as the level of decomposition increases; therefore, the wavelet packet tree can be considered as a bank of filters with each component considered as a filtered component within the bank. A good time resolution comes at the expense of low frequency resolution at the top of the WP component tree, whereas a good frequency resolution can be found at the base of the tree. Thus, the frequency resolution of the deconstructed component with high frequency components can be improved by using wavelet packet analysis. With the wavelet packet analysis, this research has more control over the signal's decomposition's frequency resolution.

There is a function for each wavelet packet, $\psi_{j,k}^i$, where 'i' is the module parameter, 'j' is the dilation parameter and 'k' is the translation parameter.

$$\psi_{j,k}^i(t) = 2^{-\frac{j}{2}}\psi^i(2^{-j}t - k) \quad (2.9)$$

Here $i = 1, 2 \dots j^n$ and 'n' is the wavelet packet tree decomposition level. The wavelet ψ^i is through the following recursive affiliations:

$$\psi^{2i}(t) = \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} h(k) \psi^t \left(\frac{t}{2} - k \right) \quad (2.10)$$

$$\psi^{2i+1}(t) = \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} g(k) \psi^t \left(\frac{t}{2} - k \right) \quad (2.11)$$

Here $\psi^i(t)$ in the discrete filters and as a mother wavelet $h(k)$ and $g(k)$ are scaling and the mother wavelet function use quadrature mirror filters.

The coefficients of wavelet packet, $c_{j,k}^i$ corresponding to the function $f(t)$ can be expressed as,

$$c_{j,k}^i(t) = \int_{-\infty}^{\infty} f(t) \psi_{j,k}^i(t) dt \quad (2.12)$$

If the orthogonality criteria is met for the wavelet coefficients, then yes.

It is possible to retrieve the signal's wavelet packet component at a specific node as

$$f_j^i(t) = \sum_{k=-\infty}^{\infty} c_{j,k}^i \psi_{j,k}^i(t) dt \quad (2.13)$$

Following the decomposition of a wavelet packet up to j^{th} level, Wavelet packet summing can be used to represent the original signal j^{th} . Equation 4.6 shows that this is the case.

$$f(t) = \sum_{i=1}^{2^j} f_j^i(t) \quad (2.14)$$

Scaling coefficients (high and low pass branch in the tree structure) of the current level are divided by sorting and downsampling in the DWT decomposition to produce the next level coefficients. The highpass branch of the binary tree (filtering and downsampling) of the wavelet packet decomposition is likewise broken up by filtering and downsampling.

2.6.2 Convolution Neural Network (CNN)

Convolution Neural Network (CNN) is a special kind of ANN (Artificial neural network) that uses convolution operation in at least one hidden layer. They are successfully used for different tasks involving processing data with grid-based structure, especially images. Figure 2.6 depicts the speech enhancement using CNN, where x denotes noisy speech signal and $f(x)$ denotes denoised speech signal.

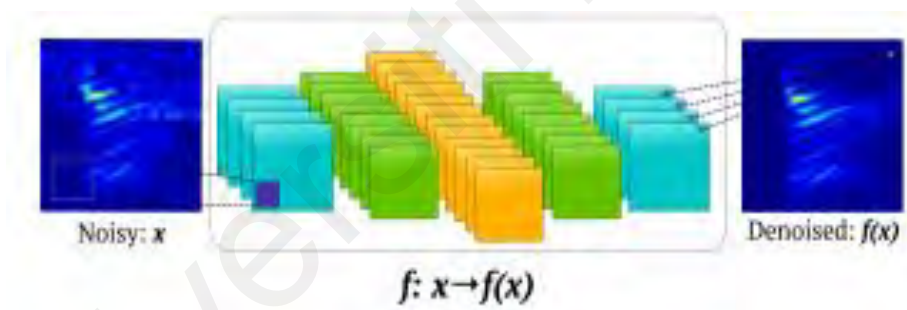


Figure 2.6: Architecture of CNN in speech enhancement (Se Rim et al., 2016)

The components of CNN layer are as follows:

(a) Convolutional Layer

The original image or another feature map can be utilised as the input for CNN's convolutional method to create a feature map from it. Convolution in ANNs is mostly used to rely on the unique structure of the input and learn how to change it into the most informative form. Convolutional

layer behaviour is controlled by a collection of input parameters, enabling neural network flexibility in design and enabling it to be adapted to different challenges.

- The convolution kernel's dimensions are defined by the kernel size. The input region that neurons are sensitive to is controlled by it. The dataset almost always determines the best value to choose for this parameter. To capture key information, such as edges, one approach is to adjust the first layer's kernel shape according to the scale of the images. For deeper layers, there isn't a standard rule, and the ideal kernel size is established experimentally. The kernel itself is frequently three-dimensional when the input consists of multi - channel images or any other three-dimensional data.
- Since each kernel would produce a unique feature map, the number of kernels affects the number of dimensions for the layer output. In architectures where the size of the feature map tends to decrease with each layer, increasing the number of kernels might help to reduce information loss. It also regulates the model's capability because the total number of trainable parameters increases with the number of kernels.
- *Padding*: Since some of the kernel cannot be matched with any input value, convolution is undefinable at the input's edges. Input can be framed with zeros in order to get around this issue and use convolution in such situations. Padding can be used to control the output size because the size of the output directly depends on the number of input values for which the convolution is defined.
- The convolution filter's step is controlled by stride. A stride value of two indicates that two pixels are skipped across all dimensions once convolution is applied to a given pixel. This research can control how different receptive fields overlap and how much output is produced by adjusting stride. Let n_{in} , n_{out} , k , p , s represent the sum of inputs and outputs,

respectively, as well as the total kernel size, padding size, and stride. Consequently, the relationship shown below is valid:

$$\text{iii. } n_{out} = \left\lfloor \frac{n_{in} 2^{p-k}}{s} \right\rfloor + 1 \quad (2.1)$$

(b) Pooling

The three phases of a typical convolutional layer are as follows:

- Using convolutions to achieve an intermediate outcome.
- Passing intermediate outcome through a non-linear activation function, similar to that of the standard multilayer perceptron's. It is also known as the detecting stage.
- Apply pooling function

The rectangular input regions are replaced with their summaries by the pooling procedure. It can be seen of as a non-linear way of down sampling.

Pooling is an important subject that transforms the combined feature representation into critical data by retaining helpful information and removing irrelevant information. Pooling is a useful algorithm for handling small frequency alterations that are typical in speech signals. Additionally, pooling helps to reduce the spectral variance in the input speech. It applies a particular function to transform the input from p neighboring units into the output. The pooling layer receives the features after they have been subjected to the element-wise non-linearities. The previous layer's feature maps are down-sampled in this layer, resulting in new feature maps with a reduced resolution. The input's spatial dimension is greatly reduced to this layer (Abdel-Hamid et al., 2012). Using it accomplishes both goals at once. One of the first benefits is a 65 percent

reduction in the number of parameters or weights. Secondly, it prevents the training data from becoming overfit. When a model becomes overly dependent on the training data, it is said to be overfit. Pooling methods that have been successful in computer vision tasks are examined for speech recognition tasks in this section. When the preceding convolutional layer feeds inputs to the pooling layer, it downsamples the inputs to get a single output from that region. CNNs use max pooling as their primary approach. The pooling region's maximum value is selected. Equation 2.2 contains the formula for maximum pooling.

$$s_j = \max_{i \in R_j} a_i \quad (2.2)$$

where R_j is a pooling area and $\{a_1, \dots, a_{|R_j|}\}$ is a set of functions. With max pooling, prediction accuracy of the training data is a serious issue, according to Zeiler and Fergus (2013). Stochastic and Lp pooling are two other methods for dealing with this issue. Lp pooling, according to (Bruna et al. 2013) provides better generality than maximum pooling, according to the authors.

Weighted averages are taken in the pooling region in Lp pooling. Lp pooling is depicted in Equation (2.3).

$$s_j = \left(\sum_{i \in R_j} a_i^p \right)^{1/p} \quad (2.3)$$

If $p = 1$ thus the pooling works like an average, while $p = \infty$ leads to maximum pooling. All components in the pooling zone are investigated and their average is calculated, areas of high engagement are downweighted by areas with low activation. In the case of average pooling, this is by far the most problematic issue. Stochastic pooling is a pooling approach that addresses the

max and average pooling concerns. An initial step in stochastic pooling is to normalize activations in each region R_j to get the probability p .

$$p_i = a_i / \sum_{k \in R_j} a_k,$$

$$s_j = a_l \text{ where } l \sim P(p_1, p_2, \dots, p_{|R_j|}) \quad (2.4)$$

Using the multinomial distribution created by these probabilities, the location l and the pooled activation associated with it are chosen for a_l . Multinomial distribution probabilities are used to generate the activations, which are then chosen at random. Due to the stochastic nature of stochastic pooling, overfitting is not possible. Stochastic pooling has the same advantages as maximum pooling.

(c) *Batch Normalization*

The output of the layer before the batch normalization layer is normalized. If $X \subseteq$ is a collection of inputs, then:

Algorithm: compute the output y of the batch normalization layer

$$\mu \leftarrow \frac{1}{|X|} \sum_{x \in X} x$$

$$\sigma^2 \leftarrow \frac{1}{|X|} \sum_{x \in X} (x - \mu)^2$$

$$\hat{x} \leftarrow \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

$$y = Y\hat{x} + \beta$$

Batch normalisation enhances the gradient computation by centering and scaling the feature maps. The layer's primary objective is to eliminate any unwanted training-related changes to the hidden layer output distribution. As a result, learning is made easier and the convergence to the minimum is faster. In the process of learning, $\gamma \in \mathbb{R}$ and $\beta \in \mathbb{R}$ values are established.

(d) Non-Linear Units and Dropout

Sigmoid, maxout, rectified linear units (ReLU), and proportional rectified linear units (PReLU) are some of the non-linear functions that can be evaluated in this section by altering the type of completely linked layers.

- Sigmoid Neurons

The typical sigmoid function is the best fit for acoustic modelling. Its advantages lie in the fact that there are no variable parameters. Many applications of the sigmoid family in ASR have been investigated (Saon et al., 2014).

$$f(\alpha) = \eta \frac{1}{1+e^{-\gamma\alpha+\theta}} \quad (2.5)$$

$f(\alpha)$ is the function, and θ , γ , and η , the learnable parameters are referred to as such. The p-sigmoid can be found in equation (2.5) (η , γ , θ).

- Relu

Two forms of neural network training are commonly used. To begin, CNN is trained using a frame discriminative SGD CE (Cross Entropy) criterion. For the second time, the CE-trained CNN weights are re-adjusted using the sequence level objective function. Speech recognition is more of a non-stationary sequence task; hence the second type is more significant.

In comparison to a CE-trained CNN, several studies have shown that sequence training boosts ASR performance by 10–15%. Sequence training can benefit from second-order HF (Hessian-free) optimization, although this is less critical for CE training. Using ReLUs and dropouts, Srivastava et al. 2014 developed a new algorithm to regularize CNNs. Dahl et al. 2021 have shown that CE trained CNNs utilizing ReLU + dropout achieved a 5% relative reduction in WER on a 50-hour English Broadcast News large vocabulary ongoing speech recognition (Sainath et al.,2013). ReLU is a linear activation function that is not saturated. For negative numbers, the output is 0 and the input is itself. reLU is expressed as:

$$h_l = \max(0, z_l) \quad (2.6)$$

CNN training without dropout yielded significant increases and performance, but further HF sequence training without dropout wiped off some of those gains and results. For example, dropout is effectively used with HF sequence training in this study.

- Parameterized Rectified Linear Units

When the ReLU units are inactive, they produce no gradients. Consequently, they will not have their weights updated using gradient-based optimization. The training process is slowed down because of the use of constant zero gradients. Scholars proposed PReLU, an enhanced version of ReLU that incorporates the negative component to speed up learning, to address this problem (K. He et al., 2015). Thus, the vanishing gradient problem is solved satisfactorily. If the input is negative, then the output is multiplied by a slope of α ; otherwise, the input is multiplied by its own value. This is how a PReLU function is defined:

$$h_l = \begin{cases} \alpha z_l & \text{if } z_l < 0 \\ z_l & \text{otherwise} \end{cases} \quad (2.7)$$

- *Dropout*

The under-fitting problem is expertly handled by Maxout neurons, resulting in improved optimization efficiency (Toth et al., 2014). Due to their huge capacity, CNNs adopting maxout non-linearity are particularly susceptible to overfitting. Regularization approaches such as Lp-norm, weight decay, weight tying, etc., have been developed to alleviate the overfitting problem. Regularization algorithm dropout, developed by Srivastava et al. (2014), showed promise in reducing overfitting. For each training sample, half of the activations in a layer are randomly set to zero. This prevents the hidden units from co-adapting to each other and learning improved representations for the inputs, which is detrimental to the system. Dropout, as demonstrated by (Goodfellow et al. 2013), is a successful method for reducing overfitting in maxout networks due to improved model averaging. Dropout regularization is accomplished using a variety of methods during the training and testing phases. According to dropout's feed-forward operation, neural networks are trained by discarding each hidden unit at random. By isolating them from each other, this allows for sophisticated co-adaptations between hidden units to be anticipated. Especially by using dropout, y^l is given in Equation (2.8):

$$y^l = f\left(\frac{1}{1-p} \cdot W^l(r^{l-1} * y^{l-1}) + b^l\right) \quad (2.8)$$

where y^l is the layer's activation l . y^{l-1} is the layer's input l . W^l layer's weight matrix is shown in l . b^l represents a layer-specific bias vector l . $f(\cdot)$ the sigmoid or maxout activation function, respectively, and may be used to create a binary mask with a probability distribution of Bernoulli. Each entry is drawn as if it were 1 in length. Dropout is not used while decoding.

It is the ratio of the number of neurons to be removed from training to improve generalization that is referred to as the hyper-parameter p . Higher p values suggest more aggressive

regularization, while lower p values refer to more information. Throughout the course of the examinations, the factor $\frac{1}{(1-p)}$ is employed to ensure that no units are dropped during the testing process and that the entire input is sent to all the subsequent layers. Training with a dropout factor is more effective in which $(1 - p)$ is utilized to reduce the number of neuronal firings. An intelligent approach to model averaging and generalization can be found here.

(e) Fully connected layer

All learning architectures, Fully Connected Layer (FCN) plays an important role. All of the inputs from one layer are connected to each activation unit of the following layer in this architecture.

(f) Softmax layer

Soft max is the final layer of the proposed deep learning which assigns the decimal probabilities to each class in a multi-class problem.

2.7 Performance Evaluation methods for Speech Enhancement Algorithms

The performance of speech enhancement algorithms can be measured with different parameters: output Signal to Noise Ratio, stability, complexity, convergence rate, speed, mean square error, overall quality, background noise distortion, and signal distortion.

- ***Output SNR (signal to noise ratio)***

It is the most widely accepted and well-liked method for evaluating speech quality. The signal-to-noise power ratio is expressed in decibels (Pinki et al.,2015).

- **Word error rate (WER) and word recognition rate (WRR)**

For the purposes of assessing the Multi-Channel Speech Enhancement systems, the word error rate (WER) was utilized to calculate the word recognition rate (Mokbel et al., 1996).

- $$WER = \frac{S+D+I}{N} \quad (2.15)$$

Where N is the sentence's total word/letter count; S is the number of words substituted; and D is the number of words deleted. In a sentence, it represents the total number of insertions.

The word recognition rate (WRR) can be calculated using the word error ratio (WER) (Mokbel et al., 1996) :

- $$WRR = 1 - WER \quad (2.16)$$

WRR indicates the performance accuracy of Multi-Channel Speech Enhancement system.

- **Stability**

The nature of the system that generates output is determined by stability. MSE (mean square error) reduction results in decreased feedback for system stability (Anand Krishna B et al., 2016).

- **Complexity**

In order to get a quick response, complexity is measured in terms of the amount of computations, such as adders and multiplications, and it is kept as low as feasible (Anand Krishna B et al., 2016).

- **Convergence rate**

The speed of responsiveness and filtering quality are determined by the convergence rate. Convergence rate is the adjustment factor used to correct the filter after receiving feedback (Anand Krishna B et al., 2016).

- ***Speed***

The system feedback's properties are determined by speed (Anand Krishna B et al., 2016).

- ***Mean square error***

Another metric that is traditionally used to assess the degree of similarity between signals is the Mean Squared Error (MSE) (Pinki et al., 2015). The quality of the response is commonly defined by MSE.

- ***Speech Signal distortion, Background noise distortion and Overall Quality***

They should base their assessments of overall quality on the speech signal, the environmental noise, or both. This technique tells the listener to pay attention to and rate the improved speech signal in turn (Hu and Loizou 2006).

- PESQ (perceptual evaluation of speech quality): ITU-T advises using the PESQ measurement, a complex parameter, to evaluate speech quality. The average asymmetrical disturbance A_{ind} and average disturbance D_{ind} are combined linearly to create the PESQ. Researchers can calculate this as:

- $$PESQ = a_0 + a_1 D_{ind} + a_2 A_{ind} \quad (2.17)$$

Where a_0 , a_1 and a_2 are the three constant parameters whose values are 4.5, -0.1 and -0.0309

- Log-likelihood ratio (LLR): the LLR is computed as

$$d_{LLR}(\vec{a}_p, \vec{a}_c) = \log \left(\frac{\vec{a}_p R_c \vec{a}_p^T}{\vec{a}_c R_c \vec{a}_c^T} \right) \quad (2.18)$$

Where \vec{a}_c denotes the LPC vector obtained from original speech frame, \vec{a}_p denotes the LPC vector obtained from the enhanced speech frame and R_c represents the autocorrelation of original speech signal.

- Itakura-Saito (IS): the IS parameter can be computed as follows:

$$d_{IS}(\vec{a}_p, \vec{a}_c) = \frac{\sigma_c^2}{\sigma_p^2} \left(\frac{\vec{a}_p R_c \vec{a}_p^T}{\vec{a}_c R_c \vec{a}_c^T} \right) + \log \left(\frac{\sigma_c^2}{\sigma_p^2} \right) - 1 \quad (2.19)$$

Where σ_c is the LPC gain of clean signal whereas and σ_p represents the LPC gains of enhanced speech signal.

- Cepstrum coefficients: the CC can be obtained as follows:

$$d_{CEP}(\vec{c}_c, \vec{c}_p) = \frac{10}{10 \log 10} \sqrt{2 \sum_{k=1}^p [c_c(k) - c_p(k)]^2} \quad (2.20)$$

Where \vec{c}_c denotes the LPC gain of clean signal and \vec{c}_p represents the LPC gains of enhanced speech signal.

- MOS (mean opinion score) MOS assigns a value to the overall quality of the delivered speech through a network in comparison with the original speech. MOS ratings have a range from 1 (bad) to 5 (excellent) (Ramana A V et al., 2012).
- STOI (Short time objective intelligibility): Intelligibility measure which is highly compared with intelligibility of degraded speech signals. STOI ratings have a range from -0.5 to 4.5 (Higher the value implies better quality) (Taal H C et al., 2011).

2.8 Summary

This chapter presented a literature overview of speech enhancement alongside speech recognition technology for speech communication devices and various other applications. The performance of speech recognition depends on the signal quality. It begins with a brief survey of the various speech features that are commonly used in the speech recognition tasks.

Universiti Malaya

CHAPTER 3: THE MULTI-CHANNEL SPEECH ENHANCEMENT SYSTEMS IN NOISY ENVIRONMENT: THE BENCHMARK EXPERIMENT

3.1 Overview

This chapter presents the experimentation of Multi-Channel Speech Enhancement systems under noisy environments for achieving objective 2 of this research which is considered as the benchmark experiment. The purpose of experimenting Multi-Channel Speech Enhancement is to find real problem existing under noisy environments at different levels of SNR. Moreover, this chapter covers experimental designs and setup, database used, and evaluation. The benchmark experiment was conducted under the following steps:

1. Evaluate the Multi-Channel Speech Enhancement system under noisy environments with regards to spectrogram analysis.
2. Evaluate the Multi-Channel Speech Enhancement system under noisy environments with regards to word recognition rate.
3. Comparison of Multi-Channel Speech Enhancement systems under stationary and non-stationary noises at various levels of SNR.

3.2 Problem Identification and the proposed solution

A narrative literature review is one of the most essential steps in every research. It helps to define the problems that require seeking solutions. Chapter 2 which includes the literature review described the state of the science related to both, DNN-based speech enhancement and Multi-Channel Speech Enhancement systems.

Conduction of the literature review or narrative literature review helps in two aspects. First is to discover the potential problem/s in Multi-Channel Speech Enhancement under noisy

environments and the second is to propose a solution for the identified problem (/s) which will help to improve the recognition rate in Multi-Channel Speech Enhancement under noisy environments.

In this research, the benchmark experiment is yet to be conducted by the existing research and these experiments are conducted to find the real problem of MCSE under noisy environments. This research aims to conduct a noise reduction experiment for a Multi-Channel Speech Enhancement (MCSE) system in stationary and non-stationary noisy environments at different SNR levels of speech signals. This research performs enhancement experiments using the existing Multi-Channel Speech Enhancement methods including the Fixed Beamforming, ANR, and VAD algorithms.

Furthermore, the experiment considers the low-level SNRs to high-level SNRs (-10dB to 20dB) using white Gaussian noise, airport noise, babble noise, car noise, exhibition noise and restaurant noise.

3.3 Dataset

In this research, Aurora speech dataset has been used to experiment with the Multi-Channel Speech Enhancement system under noisy environments at different SNR levels of speech signals. 25 utterances from the Aurora noisy dataset consisting of 13 unique male voices and 16 unique numbers of female voices have been selected. While the number of trials are different for different noise levels, a minimum of 25 samples for each dB level were ensured.

The selected noisy speech utterances incorporate five non-stationary environmental noise types: airport, babble, car, exhibition and restaurant as well as one stationary noise type which is the White Gaussian noise at seven different SNRs: -10dB, -5 dB, 0 dB, 5 dB,

10 dB, 15 dB, and 20 dB. For -10dB noisy speech signals. 25 utterances from the AURORA clean training dataset were selected, and theatrically mixed -10dB noisy signals with the clean training dataset. 42 different conditions for every 25 utterances were prepared. The experimental design of dataset used in Multi-Channel Speech Enhancement in this research is presented in Table 3.1.

Table 3.1: Dataset used for experimenting Multi-Channel Speech Enhancement.

Speech Enhancement	Speech Database	Training and test data	SNR/db	Types of Noises
Multi-Channel Speech Enhancement	Aurora	<ul style="list-style-type: none"> • 25 utterances of clean speech signals for training • 42 sets of noise mixed signals used for testing 	-10b, -5db, 0db,5db, 10db, 15db and 20db	Airport, Babble, Car, Exhibition, restaurant and White gaussian noise

3.4 Benchmark Experiment: Multi-Channel Speech Enhancement (MCSE)

The existing Multi-Channel Speech Enhancement (MCSE) system was used as a noise filtering method and tested in real-time environments at three SNR levels (i.e., 15db, 10db, and 5db) by Yaganoglu et al. (2021). The existing Multi-Channel Speech Enhancement improves the recognition accuracy at high SNR levels (83% at 60db) (Alessandro et al., 2017). MCSE is developed in the embedded platform as shown in the below figure, and the output from MCSE is applied in computing platforms by using speech to text engine (STT). The MCSE consists of a squared array microphone set, a recording component to record the speech signals, noisy signals, beamforming, adaptive noise reduction (ANR), and voice activity detection (VAD).

The architecture of MCSE, as shown in Figure 3.1, consists of two platforms: 1. Embedded platform, where the speech enhancing process is performed, and 2. Computing platform where the word recognition accuracy process is carried out (Alessandro et al., 2017).

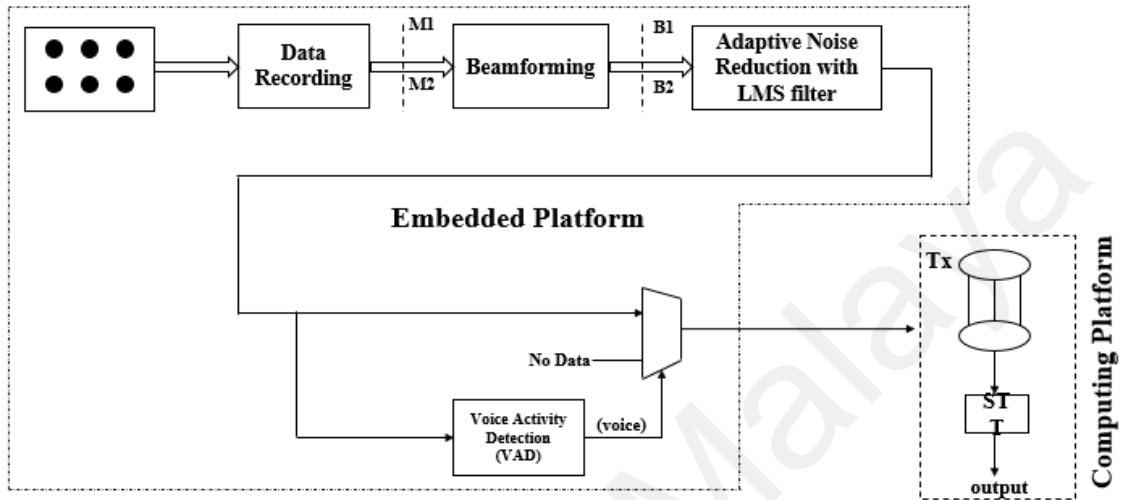


Figure 3.1: The architecture of Multi-Channel Speech Enhancement (MCSE) (Alessandro et al., 2017)



Figure 3.2: Real time architecture of Multi-Channel Speech Enhancement embedded in Helmet

3.4.1 Beamforming

There are two types of beamforming: adaptive beamforming and fixed beamforming. Adaptive beamforming is where the input signal directivity varies based on the noise signals changes in the environment. Fixed beamforming is where the input signal directivity is fixed across time, and the distance between the microphones is constant. Fixed beamforming was obtained using the delay and sum beamformer (Babu et al., 2015). Based on the microphone array in Figure 1, Alessandro et al. (2015) developed a fixed beamformer using DMA theory as explained in Alessandro et al., 2017. The output beam $B(t)$ has been calculated by applying a delay to the microphone M2 and subtracting from M1 as shown in equation (3.1)

$$B(t) = M1(t) - M2(t-T) \quad (3.1)$$

The variance in Time of Signal Arrivals between the two microphones is

$$T_0 = \frac{d}{c} \cos \theta \quad (3.2)$$

Where d is the distance between M1 and M2, c is the sound speed (constant at 20°C) and θ is the input angle. Distance between microphones and input angle is constant.

The first stream is a user beam that contains the highest SNR signals, and another stream is the noise with the lowest SNR signals. These two streams are the input signals to the adaptive noise reduction component. Fixed beamforming is not suitable for outdoor environments as the noises can come from different directions, reducing the capability of noise filtering of MCSE.

In this experiment, this research used the MEMS directional microphone array (DMA) in the squared array position to capture the signals recorded and stored in the recording component and the recommended size of the microphone is 2.7 x 1.6 x 0.89mm (L x W

x H). These recorded signals become the input signals to the beamforming component (Alessandro et al., 2017).

The beamforming device was placed at 10mm to 15mm from the left channel speaker of the sample utterance drive, while the noise generator driver was at the left corner. Python was used to write the code for mixing the speech and noise samples. Consequently, C programming language was used to write the beamformer code.

Beamformer Matlab code:

```
def multi_channel_read(prefix=r'./sample_data/20G_20GO010I_STR.CH{}.wav',
                      channel_index_vector=np.array([1, 2, 3, 4, 5, 6])):
    wav, _ = sf.read(prefix.replace('{}', str(channel_index_vector[0])), dtype='float32')
    wav_multi = np.zeros((len(wav), len(channel_index_vector)), dtype=np.float32)
    wav_multi[:, 0] = wav
    for i in range(1, len(channel_index_vector)):
        wav_multi[:, i] = sf.read(prefix.replace('{}', str(channel_index_vector[i])), dtype='float32')[0]
    return wav_multi
```

This research only tested the existing fixed beamforming algorithms in the Multi-Channel Speech Enhancement by using stationary and non-stationary environmental noises, as adaptive beamforming was never implemented on MCSE. As such, it provides the opportunity to improve the word recognition rate (WRR) accuracy rate in non-stationary environments using adaptive beamforming.

3.4.2 Adaptive Noise Reduction (ANR)

Adaptive noise reduction (ANR) is used as a multi delay block frequency adaptive filter to delete the environmental noises using an LMS filter (Soo et al., 1990; Valin et al., 2007). User beam and reference noise are the inputs to the ANR. The ANR component filters the noise in the user beam that is consistent with reference noise (speech signal already exists in the user beam as beamforming and not attenuated) (Widrow et al., 1975).

In general, assumptions cannot be made that noise was filtered in a speech signal. In this scenario, the LMS filter used by the ANR partly suppresses and changes the required signal, which depends on the attenuation of the speech signal in the reference beam and user beams. The SNR of the output signal is defined in (Palla Alessandro et al., 2017).

$$SNR_{output} = \frac{1}{SNR_{rfn}} \quad (3.3)$$

Where SNR_{rfn} is Signal-to-Noise ratio (SNR) of reference noisesignals.

The output signals of the beamforming device are the input signals of adaptive noise reduction. The LMS algorithm was applied and written in python to filter the noise signals using the MATLAB simulations and ARM processor, which filters noise in real-time.

```
self.sampling_rate = fs
    self.spec_average = (self.frame_size)/(self.sampling_rate)
    self.beta0 = (2.0*self.frame_size)/self.sampling_rate
    self.beta_max = (.5*self.frame_size)/self.sampling_rate
    self.leak_estimate = 0
```

3.4.3 Voice Activity Detection (VAD)

The Voice Activity Detection (VAD) algorithm separates the user's voice in the user stream (Alessandro et al., 2017), which is helpful for two reasons:

- (i) *Segmentation*: the system needs to know the exact boundaries of each word in the spoken utterance.
- (ii) *Data Reduction*: the system only sends data as required and not continuously over the transmission channel.

The VAD is implemented in a time domain, and the recognition is performed every 20 seconds, taking 20 samples per frame to calculate the Zero crossing and Energy

characteristics using the angular windows frame function (Venkatesha Prasad et al., 2002) (Bachu et al., 2008).

$$\sum_0^n x^2(n) \text{Energy} = \quad (3.4)$$

$$\text{Zero Crossing} = \sum_0^n \text{sign}(x[n] - \text{sign}(x[n - 1])) \quad (3.5)$$

Where $x[n]$ indicates the number of samples per frame. When Zero Crossing is small and Energy is high, it is categorized as voiced speech signal, otherwise it is deemed to be the unvoiced region of the speech signal.

In discrete time signal processing, zero-crossing occurs if the successive samples of the signal have different algebraic signs. The zero-crossing is a measure of the frequency content of a signal, i.e., the rate at which ZC occurs is the measurement of the frequency content of the input signal. It provides the total count in each time interval that the amplitude of the speech signal passes through the value of zero and can be expressed as

$$\text{sign}[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases} \quad (3.6)$$

The output audio signals from the adaptive noise reduction will be the input signals for voice activity detection (VAD). The VAD algorithm, written in python language, identifies the presence or absence of speech in audio signals.

```
def _calculate_normalized_energy(self, data):
    data_freq = self._calculate_frequencies(data)
    data_energy = self._calculate_energy(data)
    #data_energy = self._znormalize_energy(data_energy) #znorm brings worse results
    energy_freq = self._connect_energy_with_frequencies(data_freq, data_energy)
    return energy_freq
```

3.5. Experimental Design and Setup

3.5.1 Experimental design for Multi-Channel Speech Enhancement

This research aims to conduct a noise reduction experiment for a Multi-Channel Speech Enhancement (MCSE) system in stationary and non-stationary noisy environments at different SNR levels of speech signals. The research performed several enhancement experiments using the existing Multi-Channel Speech Enhancement methods which include the Fixed Beamforming, ANR, and VAD algorithms. The experiment intends to examine the MCSE system's performance for stationary and non-stationary environmental noises. This experiment considers the low-level SNRs to high-level SNRs (-10dB to 20dB) using white Gaussian noise comprising, non-stationary noise and airport noise, babble noise, car noise, exhibition, and restaurant – all categorised under stationary noises as presented in Table 3.3. The procedure of this experiment begins with the development of Multi-Channel Speech Enhancement (MCSE) based on Beamforming, Adaptive Noise Reduction, and Voice Activity Detection algorithms, as presented in Table 3.2, with the detailed procedure explained in the section below.

Table 3.2: Experimenting the Multi-Channel Speech Enhancement (MCSE) includes Beamforming, Adaptive noise reduction and Voice activity detection algorithms Using Different Types of Environmental Noises at Different SNR Levels

Algorithms	Environment	Types of noise	SNR levels (dB)
Fixed Beamforming, Adaptive Noise Reduction, Voice Activity Detection	Stationary Noises	White Gaussian Noise	-10dB, -5dB, 0dB, 5dB, 10dB, 15dB, 20 Db
	Non-Stationary Noises	Airport Babble Car Exhibition Restaurant	-10dB, -5dB, 0dB, 5dB, 10dB, 15 dB, 20 dB

3.5.2 Experimental Setup for Multi-Channel Speech Enhancement

Hardware requirements in this experiment is given below:

1) Device Configuration

- (a) Controller: stm32f103CBT6
- (b) Main clock: 72MHz
- (c) Memory: 128KB ROM/ 20KB RAM
- (d) External storage: Transcend 8GB class 4 memory card.
- (e) Transducers: MEMS microphones (Bandwidth 20KHz., Capacitance 1 pf, Bias voltage 10v, Capacitance variation 10 fF/Pa, Sensitivity 100 mV/Pa).

2) Sampling Setup

Two transducers' output was pre-amplified, then fed to a single-stage bandpass filter(80Hz-16KHz), and later gain adjusted, level shifted to 1.75V, and was thereafter fed to individual ADC's (analog to digital converter).

ADCs at 12bit vertical resolution and 16000 Samples per second was configured (+/- 50 due to clock stability). Data was written to SD card via conversion and had a complete interrupt linked to DMA channel which wrote the value in SD card and a copy in Buffer variable defined in RAM. Both ADC sampling times were synchronized. Amplifiers used were based on LM358 general purpose Opamp.

3) Variability Setup

Timer 1 PWM (Pulse width modulation controller) channels connect the two 9G servos at 16bit resolution (Effective usable steps were around 30000 per servo due to higher ARM deflection of Servos). The distance between each microphone is fixed at 10 mm.

4) Noise and Sample Utterance System Setup

The primary noise driver is Edifier 2.0 channel speaker. The speech is varied at the amplifier and the noise samples were continuously looped and fed to the amplifier from the BeagleBone Black Single Board. The speech samples were driven with only left channel speakers of Logitech USB speakers and the BeagleBone Black single-board computer fed the samples.

5) SNR Setup

The desired SNR (-10dB, -5dB, 0dB, 5dB, 10dB, 15dB and 20dB) was achieved by individually tuning the noise sound amplifier gain control and the sample utterance amplifier gain control by measuring individual Sound Pressure levels (SPL) to the calculated values.

3.6 Evaluation

In this research, the spectrogram analysis, and the Word Recognition Rate (WRR) were used to evaluate the performance of the Multi-Channel Speech Enhancement system in a noisy environment (stationary and non-stationary noise).

3.6.1 Spectrogram Analysis

Spectrogram analysis is used to analyse the amplitude of speech signals (Haykin et al., 1991). A spectrogram analysis shows the spectral illustrations of a time-varying signal (Flanagan et al., 1972). The spectrogram analysis for both stationary (White Gaussian Noise) and non-stationary environmental noises (Babble, Airport, Car, Exhibition, and Restaurant) are performed on time-domain using MATLAB.

3.6.2 Word Error Rate (WER)

This research evaluates the voiced speech signal received after the voice activity detection with the ASR speech-to-text engine to determine the word error rate (WER).

WER is calculated to evaluate the performance of the Multi-Channel Speech Enhancement systems. WER is computed as follows:

$$\text{Word Error Rate (WER)} = \frac{\text{Insertion} + \text{Substitution} + \text{Deletion}}{\text{Number Of Words}} * 100\% \quad (3.7)$$

By calculating the WER, the word recognition rate (WRR) is determined as:

$$WRR = 1 - WER \quad (3.8)$$

WRR measures the performance accuracy of a Multi-Channel Speech Enhancement system.

3.7 Results

3.7.1 Spectrogram Analysis

Figure 3.3 represents the clean speech signal, while Figures 3.4 to 3.15 depicts the spectrogram analysis: (a) noisy speech at -5 dB SNR and (b) enhanced speech using adaptive beamforming, ANR, and VAD algorithms in MCSE.

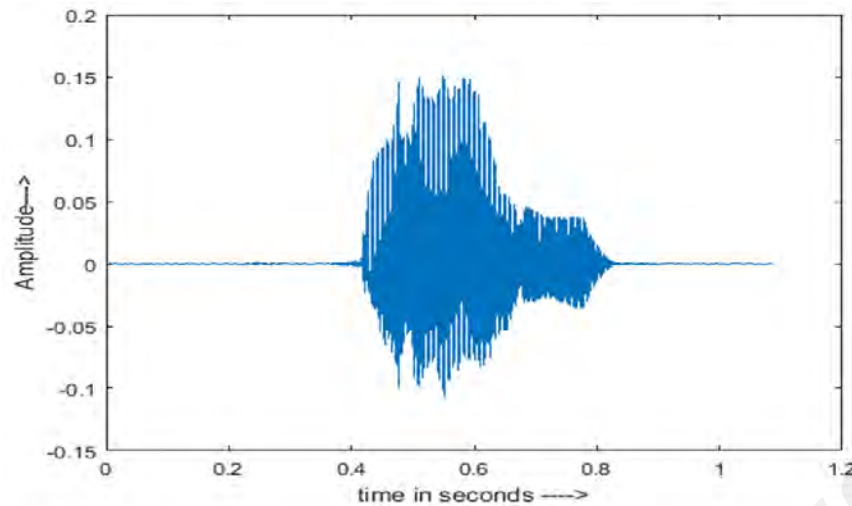


Figure 3.3: Clean speech signal

This research considered different noise types for this analysis, they include white Gaussian noise, airport noise, exhibition, restaurant, babble, and car noise. This experiment added the 5dB noise to the original signal and processed it through the considered Multi-Channel Speech Enhancement system.

The filtered signal is also known as the reconstructed signal, which can be used to analyse the performance of the existing approach.

Figure 3.4 depicts the spectrogram for unfiltered white Gaussian noise in speech signal at -5db, while Figure 3.5 shows the filtered white Gaussian noise in speech signal at -5db. From the spectrogram, speech distortion due to noise has been rectified through speech reconstruction with a suitable filter. Figure 3.6 shows the unfiltered airport noise in speech signals at -5db, while Figure 3.7 shows the filtered airport noise at -5db.

Figure 3.8 depicts the unfiltered babble noise in speech signal at -5db, while Figure 3.9 shows the filtered babble noise at -5db. Figure 3.10 shows the unfiltered car noise in speech signal at -5db, and Figure 3.11 depicts filtered car noise in speech signal at -5db. Figures 3.12 and 3.13 depict the unfiltered exhibition noise in speech signal at -5db and filtered exhibition noise at -5db, respectively. Finally, Figure 3.14 depicts the unfiltered restaurant

noise in speech signal at -5db, while Figure 3.15 shows the spectrogram for filtered restaurant noise in speech signal at -5db.

Universiti Malaya

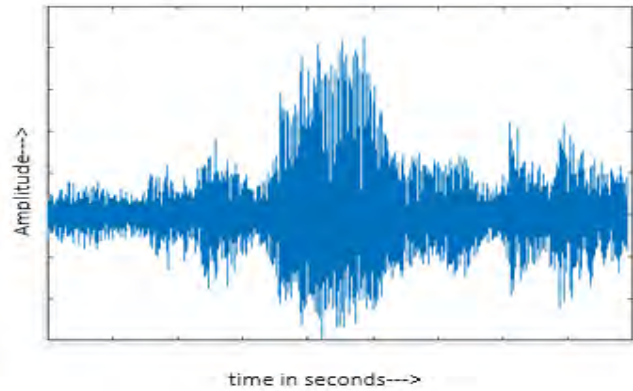


Figure 3.4: The spectrogram for unfiltered white Gaussian noise
in speech signal at -5db

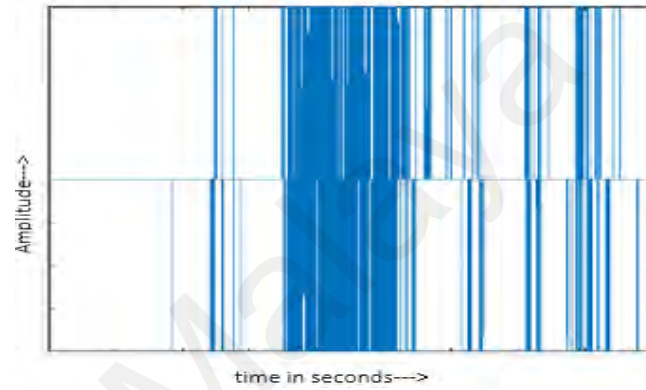


Figure 3.5: The filtered white Gaussian noise in speech signal
at -5db

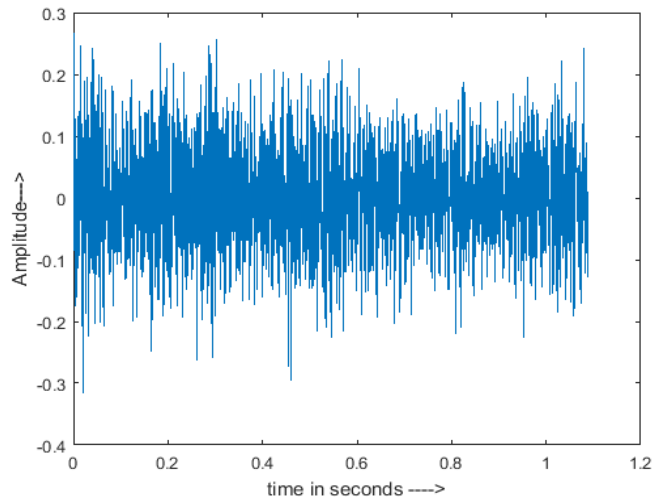


Figure 3.6: The unfiltered Airport noise in speech signal at -5dB

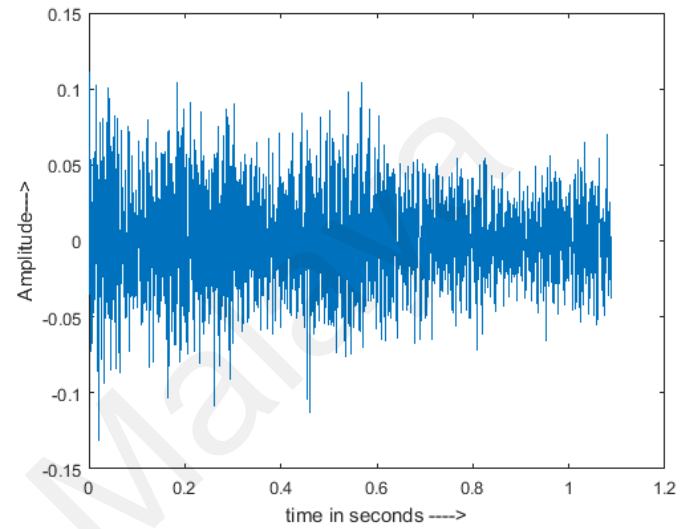


Figure 3.7: The filtered Airport noise in speech signal at -5dB.

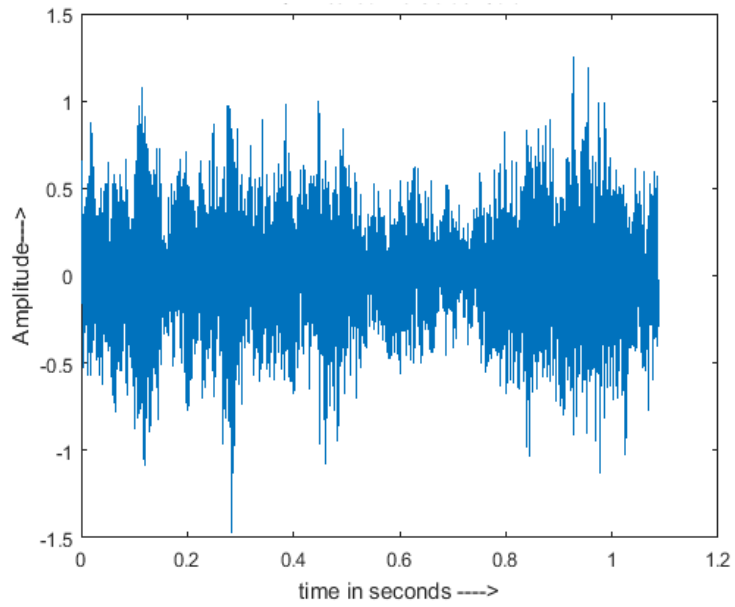


Figure 3.8: The unfiltered Babble noise in speech signal at -5dB

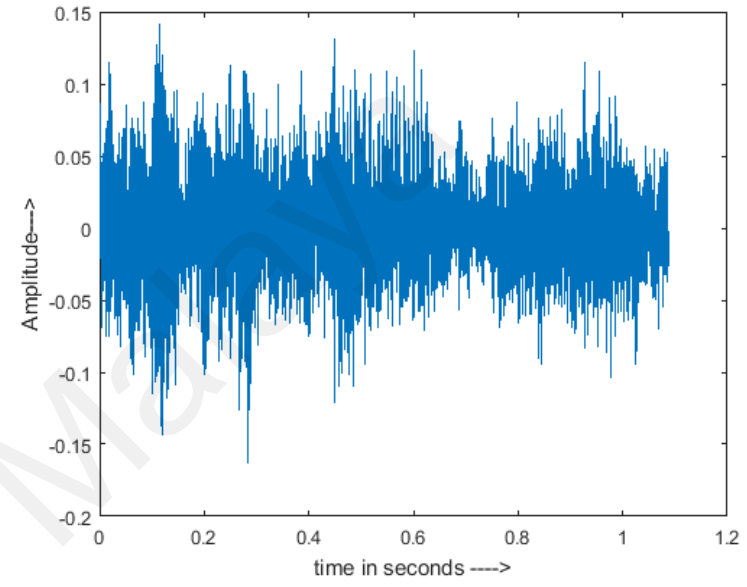


Figure 3.9 The filtered Babble noise in speech signal at -5dB

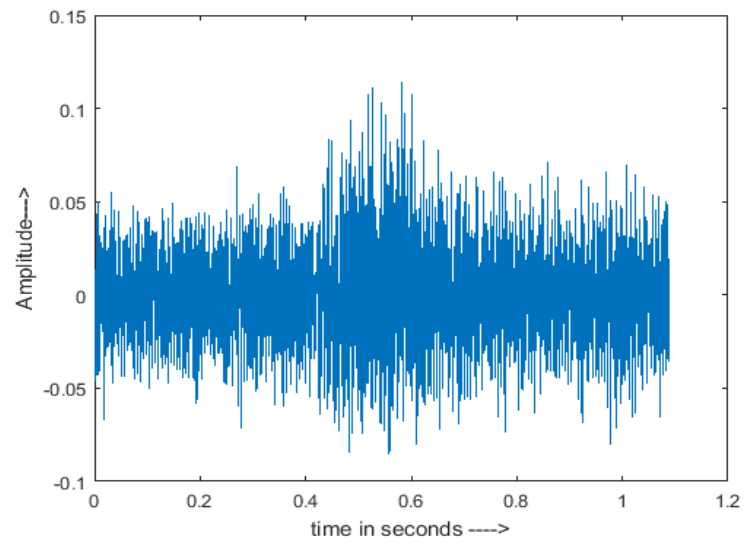


Figure 3.10: The unfiltered Car noise in speech signal at -5dB

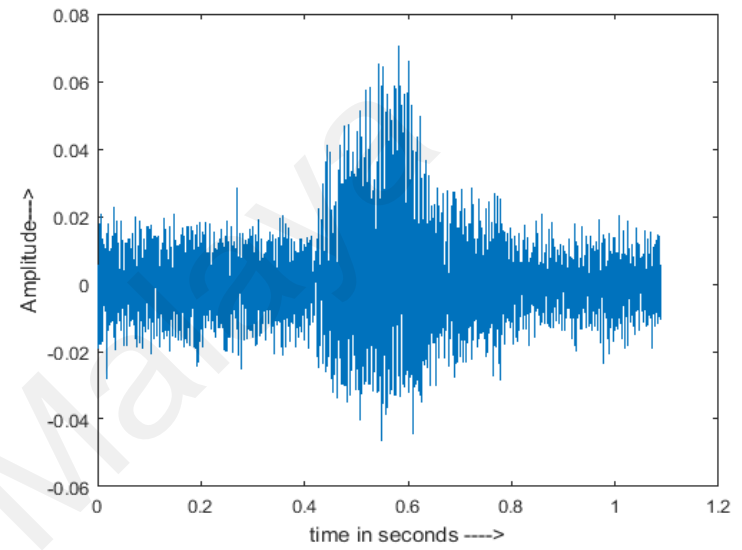


Figure 3.11: The filtered Car noise in speech signal at -5dB

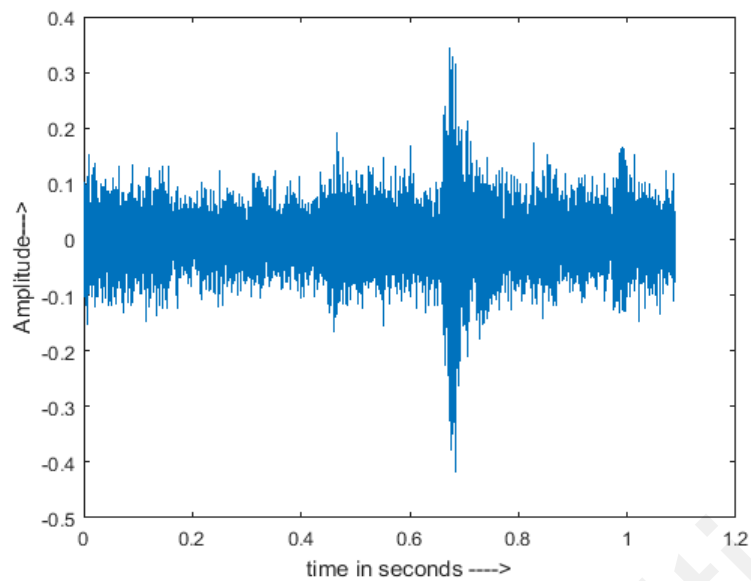


Figure 3.12: The unfiltered exhibition noise in speech signal at -5db

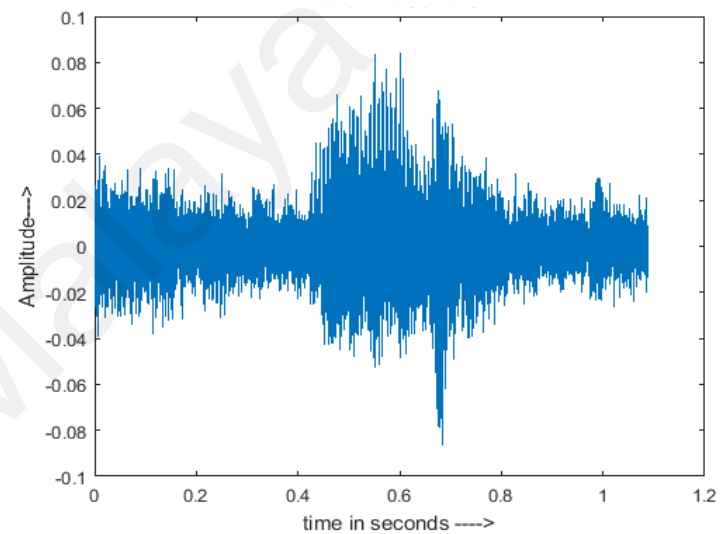


Figure 3.13: The filtered exhibition noise in speech signal at -5db

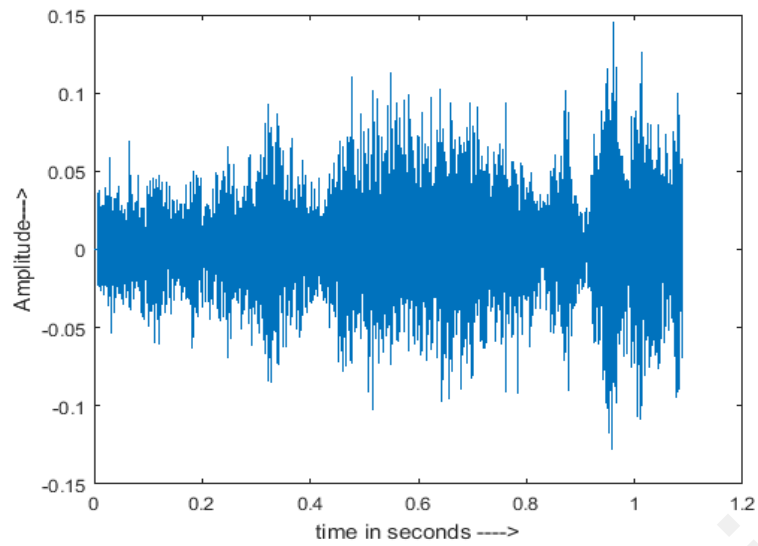


Figure 3.14: The unfiltered restaurant noise in speech signal
at -5db

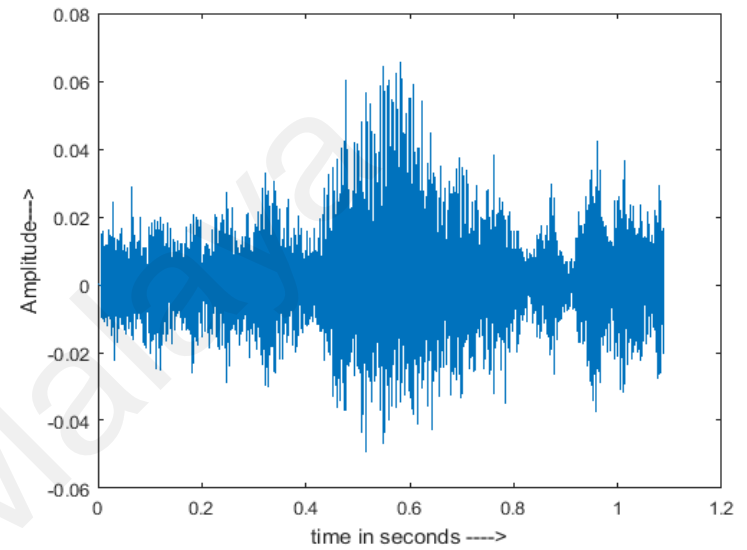


Figure 3.15: The spectrogram for filtered restaurant noise in
speech signal at -5db

3.7.2 Word Recognition Rate (WRR)

Table 3.3 and Table 3.4 presents the evaluation results of the experiments using the Beamforming, ANR, and VAD algorithms in MCSE at different levels of SNRs under stationary and non-stationary noisy environments.

Table 3.3: Word Recognition Rate (WRR) For Multi-Channel Speech Enhancement (MCSE) under Stationary Environmental Noise

	SNR/dB	WRR (MCSE)
Stationary Noise White Gaussian noise	-10dB	42.6
	-5dB	58.3
	0dB	63.4
	5dB	68.5
	10dB	72.6
	15dB	74.8
	20dB	89.7

From Table 3.3, it is revealed that the MCSE with fixed beamforming, ANR, and VAD is very effective at 20dB SNR, but the recognition accuracy gradually decreases at 15dB, 10dB, 5dB, 0dB, -5dB, -10dB, respectively for the Stationary white Gaussian noise. The WRR was lower for the negative dB than the positive dB.

The scatter diagram depicted in Figure 3.16 shows a linear relationship between the SNR level and WRR (p-value < 0.001).

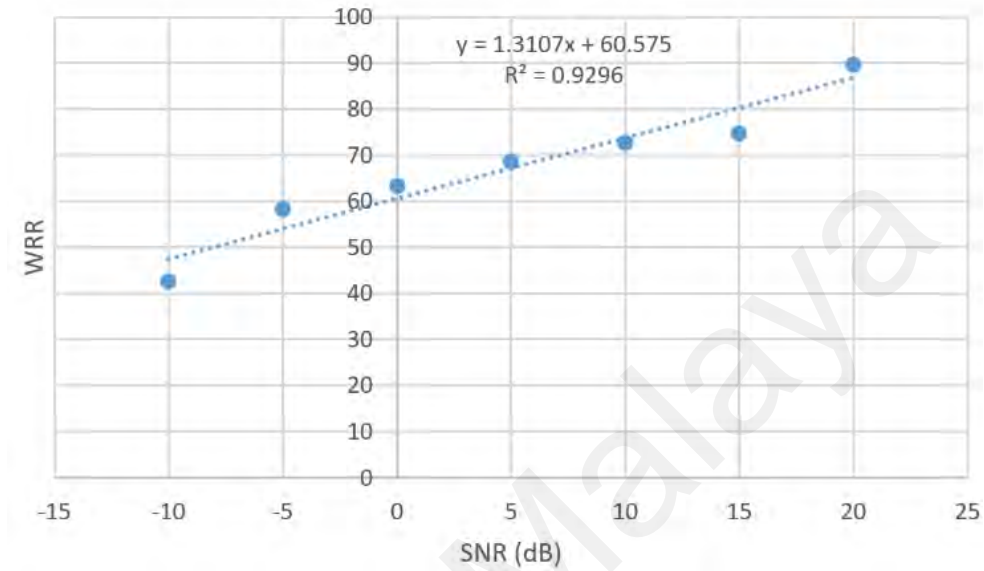


Figure 3.16: The SNR and WRR linear relationship for stationary noise in MCSE

Similarly, the result in Table 3.4 also indicates that MCSE with fixed beamforming, ANR, and VAD is very effective at 20dB SNR, but the recognition accuracy gradually decreases at 15dB, 10dB, 5dB, 0dB, -5dB, -10dB for non-stationary noise. The results show that selected methods for MCSE can better deal with environmental noise issues at 20dB, 15dB, and 10dB SNRs and can be more suitable for speech recognition applications, especially in the outdoor environment.

From Table 3.4, the average WRR for non-stationary noises was the highest for exhibition and the lowest for restaurants (a difference of about 9%). The low WRR for the restaurant is due to the mix-up of many speeches and non-speech noises.

By referring to the scatter diagram, there is a linear relationship between the SNR level and WRR. Figure 3.17 depicts the linear relationship of SNR and WRR for non-stationary noise (p-value < 0.001).

Table 3.4: Word Recognition Rate (WRR) for Multi-Channel Speech Enhancement (MCSE) under Non-stationary Environmental Noises

WRR for Non-Stationary environmental noises (%)						
SNR/dB	Airport	Babble	Car	Exhibition	Restaurant	Average
-10dB	5.82	4.04	7.26	6.54	4.54	5.64
-5dB	12.32	7.12	13.26	11.54	6.38	10.12
0dB	19.06	17.56	16.55	20.23	12.12	17.10
5dB	36.14	35	35.16	44.66	38.24	37.84
10dB	67.26	74.18	67.2	77.72	55.56	68.38
15dB	88.88	90.64	92.02	91.46	75.2	87.64
20db	93.6	91.28	97.92	94.78	91.28	93.77
<i>Average</i>	<i>46.15</i>	<i>45.69</i>	<i>47.05</i>	<i>49.56</i>	<i>40.47</i>	

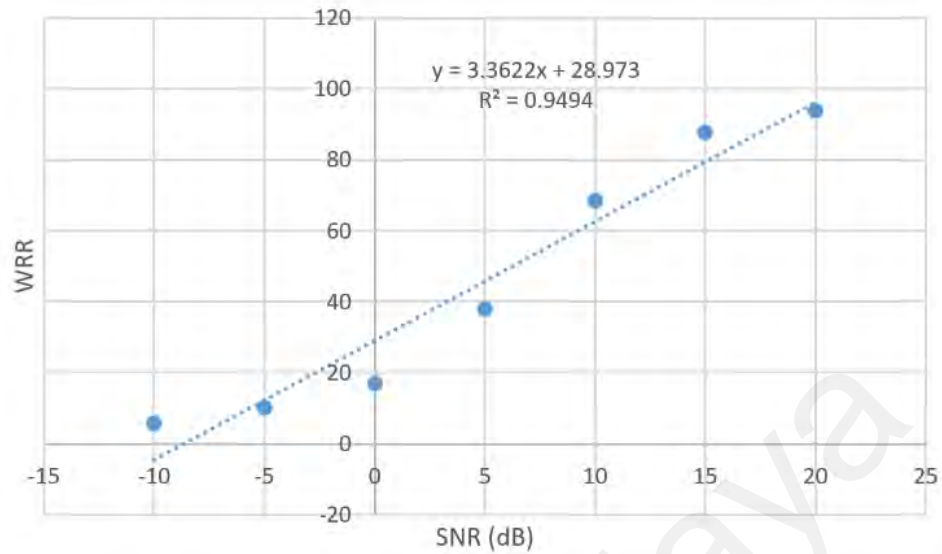


Figure 3.17: The SNR and WRR linear relationship for non-stationary noise in MCSE

When comparing the linear relationship of SNR and WRR for both the stationary and non-stationary noise, this research found that the gradient for the latter was higher than the former. It indicates that the performance of selected methods for MCSE improves at a higher rate 93.7 at 20db when SNR increases (from -10db to 20db) for non-stationary noise.

By comparing the result in Tables 3.3 and 3.4, the research found that the selected methods such as beamforming, ANR and VAD in for MCSE was better for stationary noise at low dB but was more effective for non-stationary noise at high dB noise. Figure 3.18 depicts the differences in the WRR for both the stationary and non-stationary noises for the MCSE based on fixed beamforming, ANR, and VAD. The selected methods work well for stationary noise at -10dB to 10dB. For 15dB and 20dB, the MCSE is very effective for recognizing speech in non-stationary noises.

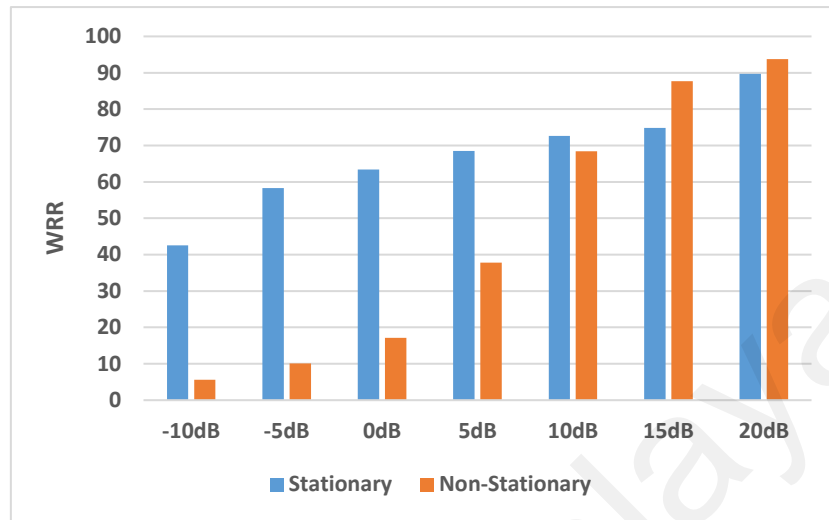


Figure 3.18: WRR for both the stationary and non-stationary noises in MCSE

Finally, to determine whether the result for stationary and non-stationary noise was significantly different, this research conducted the Analysis of Variance (ANOVA), and the result is shown in Figure 3.19 below. It was found that the linear relationship of SNR and WRR was not significantly different between the stationary and non-stationary noise. As such, the performance of the selected methods for both the stationary and non-stationary noise was statistically similar at a 95% confidence level.

ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	1594.525	1	1594.525	1.987332	0.184008	4.747225
Within Groups	9628.135	12	802.3446			
Total	11222.66	13				

Figure 3.19: Result of ANOVA for stationary and non-stationary in MCSE

3.8 Discussion

In this chapter, from the benchmark experiment, MCSE with fixed beamforming, ANR, and VAD is very effective at 20dB SNR, but the recognition accuracy gradually decreases at 15dB, 10dB, 5dB, 0dB, -5dB, -10dB, respectively for the Stationary white Gaussian noise. The WRR was lower for the negative dB than the positive dB. It also shows that MCSE with fixed beamforming, ANR, and VAD is very effective at 20dB SNR, but the recognition accuracy gradually decreases at 15dB, 10dB, 5dB, 0dB, -5dB, -10dB for non-stationary noise. The results show that selected methods for MCSE can better deal with environmental noise issues at 20dB, 15dB, and 10dB SNRs and can be more suitable for speech recognition applications, especially in the outdoor environment.

This research experimented Multi-Channel Speech Enhancement system based on fixed beamforming, ANR, and VAD algorithms and their recognition accuracy. Moreover, spectral analysis and word recognition rate evaluations were conducted on the MCSE. The word recognition rate evaluation had confirmed that the selected methods for MCSE could not perform effectively under low SNR conditions. However, it performed better in a noisy stationary environment than in non-stationary noisy environments.

This research also found that the selected methods for MCSE perform effectively at high SNR in stationary and non-stationary noisy environments. The linear relationship between the SNR and WRR has proven that the current MCSE successfully filters noise at higher SNR but fails at lower SNR (-10db, -5db, 0db). The strength of the noise is too small for the selected methods for MCSE to filter the noise away. As such, more work is needed to increase the ability of the MCSE to filter noise with low SNR, which can be a promising future direction in MCSE research.

A possible reason to get WRR with less than 80% is because of high amount noise mixed in speech signal. WRR is good at high SNR, and it is limited at low SNR'S. Therefore, this research is more focused at low level and achieved by using proposed deep learning algorithms.

3.9 Summary

In this study, the benchmark MCSE algorithms was employed. It indicates that the performance using fixed beamforming, ANR, and VAD is inferior. It suggests that the existing methods for noise reduction in MCSE, such as fixed beamforming, ANR, and VAD, are not adequate. New methods and preprocessing algorithms for noise filtration in MCSE are needed to improve the performance of the MCSE in a noisy environment.

CHAPTER 4: THE PROPOSED NOISE FILTERING FRAMEWORK FOR MULTI-CHANNEL SPEECH ENHANCEMENT SYSTEM

4.1 Overview

This chapter presents the details of the proposed approach which is developed to handle the real-time noise reduction challenges in Multi-Channel Speech Enhancement algorithms. One of the objectives of this research is to propose a framework to enhance the speech quality for speech recognition along with enhancement. From the literature review and findings from preliminary experiments in chapter 3, it has been proved that the existing MCSE performs poor with low recognition accuracy under noisy environments. This chapter presents the development of the proposed framework of Multi-Channel Speech Enhancement to improve the recognition accuracy under high to low SNR levels of environmental noises.

4.2 The Proposed Framework for Multi-Channel Speech Enhancement System

This section presents the proposed solution for speech enhancement where the research considered wavelet transform based pre-processing algorithm and deep learning based approach was also implemented. The framework of the proposed Multi-Channel Speech Enhancement is shown in Figure 4.1. below and details of each component is explained in Table 4.1.

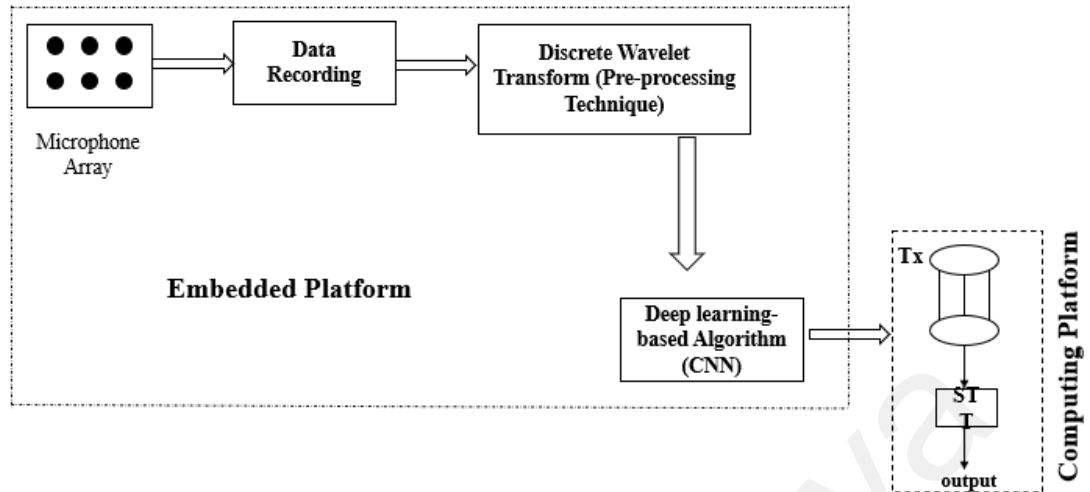


Figure 4.1: The Proposed Multi-Channel Speech Enhancement Framework based on Wavelet transform and Deep learning approach (CNN-BLSTM).

Table 4.1: Detail Explanation of each Component and their process in the proposed approach.

Module/Components	Aim	Input	Process	Output
DMA microphone array	To capture the real-time signals	Speech signals	It has two or more microphones that can be used to audio input. The array device's microphones collaborate to continuously record input signals.	Noisy speech signals
Data recoding	To store speech signals in the Hard disc	DMA microphone array data	Analog to digital converter is used to process the signal to store in hard disc	Stored data in the hard disc which is used for further process
Wavelet transform	To decompose the signal	Stored data signal	Construct the wavelet transform,	Decomposed frequency bands

		into different frequency bands		divide the frequencies into LL, HL, LH, and LL bands	
Deep learning		To learn the data patterns	Decomposed frequency band speech signals	Data is processed through multiple layers such as convolution, fully connected, pooling, max pooling etc. to learn its attributes	Enhanced learned dataset corresponding to the speech signal
Computing (ASR)	platform	To improve the recognition accuracy	Enhanced data signals obtained from Deep learning	Classify the pattern based on the ASR engine	Recognition of signals

4.2.1 Discrete Wavelet Transform

This research has adopted the Discrete Wavelet transform among other preprocessing algorithms as its performance is very effective in terms of denoising the speech signal and compressing speech signal without any significant loss in speech quality (Ping et al., 2019, Katti et al., 2011 and Maria Labied et al., 2021).

Discrete Wavelet transform is used to eliminate redundant data from an input speech signal. Wavelet transform preprocessing algorithm has been applied in input signals to obtain the temporal feature analysis (Maria Labied et al., 2021). The purpose of this algorithm is to create by rescaling and iterating through a series of filters. Up-sampling and down-sampling (subsampling) processes determine the signal's resolution (detail information), whereas filtering operations determine its scale (resolution).

From the literature review and preliminary experiments, it has been proved that MCSE performs low recognition accuracy under noisy environments. Due to lack of preprocessing implementation on MCSE system, this research used existing discrete wavelet transform preprocessing algorithm to remove the redundant data from input noisy speech signals.

Furthermore, wavelet-based preprocessing algorithm was implemented and applied on the signals obtained through speech communication devices. To compute the Wavelet Series, which is a sampled form of CWT, there may be a need for a large amount of time and resources. There is evidence that the sub-band coding-based discrete Wavelet Transform (DWT) is more efficient in computing Wavelet Transforms. It's simple to implement and decreases the amount of time and resources needed for computations. Digital filtering algorithms are used to obtain a time-scale depiction of the digital signal in DWT. Filters with various cutoff frequencies and scales are used to evaluate the input signal.

Algorithm 1: (Chiluveru et al., 2021)

Input: original noisy speech signal, wavelet decomposition bands

Output: decomposed signals and corresponding coefficients

Xdata[] stores the input data vector, and Ydata[] is the output data vector that is returned. N is the length of both data vectors. Before applying this approach, it is presumable that the wavelet filter parameters G[k] and the scale filter parameters H[k] have been provided. L is the total number of parameters. N must be an even number to work with this algorithm.

Step 1: Set $s = \frac{N}{2}$ // Start index of the input array's gamma coefficients

Step 2: Allocate ydata [N]; // Provide a memory space for the output data vector

Step 3: for (i = 0 while i < N increment i = i + 1) do // loop over input data

```

Step 4:  $ydata[i] = 0;$  // Reset summation accumulators

Step 5: endfor;

Step 6:  $j = 0;$  // access/index to the output data array

Step 7: for ( $i = 0$  while  $i < N$  increment  $i = i + 1$ ) do // loop over input
data
Step 8: for ( $k = 0$  while  $k < L$  increment  $k = k + 1$ ) do // convolution loop
Step 9:  $didx = (i + k) \bmod N;$  // access/index into input data with
wraparound.
Step 10:  $ydata[j] = ydata[i] + G[k] * xddata[didx];$  //Scaling filter
contribution
Step11:  $ydata[j + s] = ydata[i + s] + H[k] * xddata[didx];$  // Wavelet
filter contribution
Step 12: endfor;
Step 13:  $j = j + 1;$  // Update position in output array
Step 14: endfor;

```

The code which was developed in python language and implemented in matlab has been included in the Appendix VIII.

4.2.2 Deep Learning-based approach

This research has adopted the CNN among all other deep learning methods as CNN is effective in distinguishing between the speech and noise components of noisy signals because it can handle the local temporal spectral features of speech. Both in the spectrum and waveform domains, CNN has demonstrated its efficacy for improving speech. CNN approach is suitable for the proposed framework as compared to others. Additionally in this research, CNN was extended with BLSTM (Bidirectional Long Short-Term Memory) layer, because the modelling capacity of CNN is constrained. Even though convolutional neural

networks (CNNs) effectively simulate the structural locality from the feature space. Because they adopt the pooling at a limited frequency domain, they also lower the linear variance and handle disturbances and minor shifts in the feature space. By making use of prior knowledge of the speech signal, they can take advantage of the long-term dependencies between the speech frames. However, CNNs in speech communication systems cannot handle a lot of semi-clean data; as a result, the system's performance degrades. To overcome these issues, Bidirectional Long short-term memory (BiLSTM), which regulates the flow of information by an individual component called a memory block, was developed. CNN-BiLSTM is explained below in detail.

The fundamental purpose of CNN, the advanced form of DNN is to detect local structure in input data. The spectrum correlations in acoustic features are well-modelled by CNN, which successfully decreases the spectral fluctuations. Three distinct models comprising CNN, BLSTM, and fully connected layers are included in the suggested architecture as illustrated in Figure 4.3.

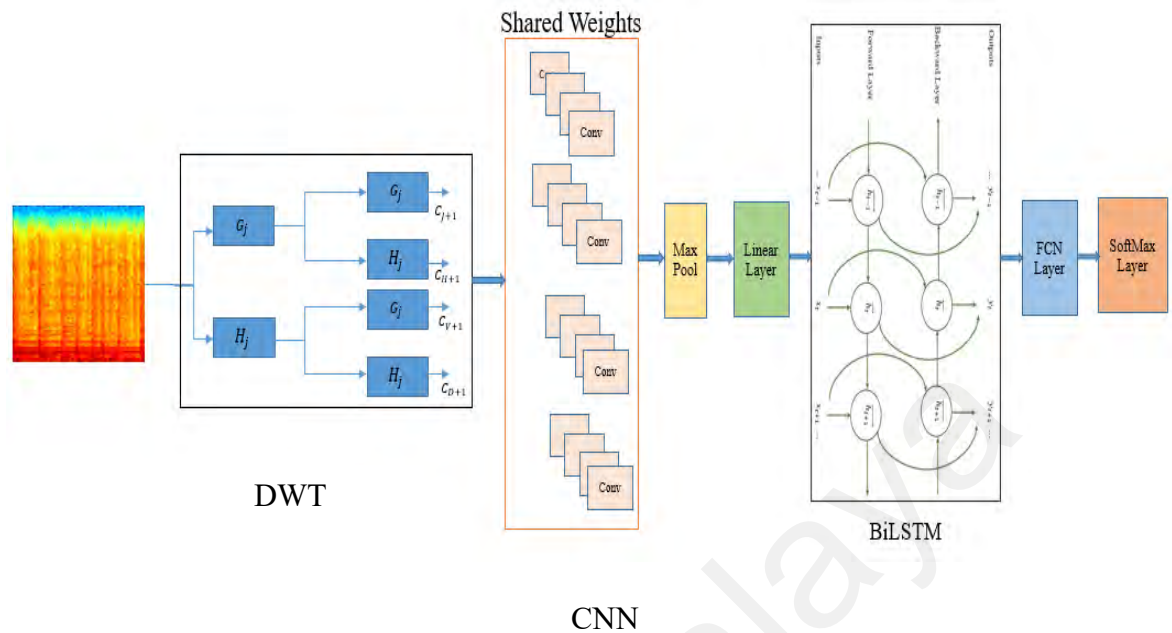


Figure 4.2: CNN-BLSTM Architecture applied for speech enhancement

Convolutional layers are used to reduce the frequency variance in the input signal initially. Two CNN layers with 256 feature mappings in each convolutional layer were chosen at first instance. This is because speech has a very tiny feature dimension (i.e., 40). The behaviour of the high- and low-frequency zones is vastly different. Nearly 16 percent of the feature map's original size has been decreased using two convolutional layers. Thus, modelling locality and eliminating invariance are no longer necessary. As stated by Sainath et al. (2015), the first convolutional layer has a 9 x 9 frequency-time filter while the second layer has a 4 x 3 frequency-time filter. A 9 by 9 frequency-time filter is used in the first convolution layer, and a 4 by 3 frequency-time filter is used in the second. In the beginning, our framework employs solely frequency-domain pooling using max pooling. Similarly, the pooling size is 2 for both layers, and the stride value is 2. The next layer in CNN has a greater dimension since the set of feature maps, time, and frequency are proportional to the layer's size. Therefore, the feature dimensions must be reduced. As demonstrated in Figure 4.2, after CNN layers, a linear layer is applied to reduce the layer's size without sacrificing accuracy.

Frequency modelling is an algorithm for reducing the dimensionality of data by using the linear layer's 236 suitable outputs. To simulate the signal in time, the output of the CNN layer is passed on to the BLSTM layer. In this case, two BLSTM and three FC layers are ideal, however the number of layers can vary depending on the experiment. Each BLSTM layer has 832 cells and 512 units (256 LSTM units per direction) of projection layer for feature extraction (256 LSTM units). Twenty-time steps are pre-trained into the BLSTM and backpropagation is truncated. The output of BLSTM layers is sent to FC layers after frequency and timing modelling. Higher-order feature representations that are easily distinguishable between classes can be generated by using these layers. A total of 1024 hidden units can be found in all fully connected layers.

Variability in speech is a result of the accent, volume, and other characteristics that distinguish distinct speakers. The proposed approach uses shared weights which are obtained by applying several convolution operations. These convolutions generate features and are supplied to the Max pooling layer. The shared weights mechanism helps to retain the top level and low-level attributes which leads to improving the accuracy. Further, these attributes are processed through the Linear Layer which supplies these features to CNN-BiLSTM layer (ermanet et al., 2012). In most CNN work, FC layers are used to discriminate between classes based on local knowledge. CNN-BLSTM module is used for energy and timing modelling, and the softmax layer is utilized to distinguish between different classes. The entire model is trained at the same time and CNN needs to retrain whenever input data is updated with 80% of ratio.

Algorithm 2: (Chiluveru et al., 2021 ; Dong wang et al., 2019)

Input: speech signals, Deep learning parameter (batch size, feature dimension, classes, train test ratio).

Output: enhanced speech signal with recognition rate performance

Step 1: capture speech signals by using DMA microphone array.

Step 2: Apply an analogue to digital converter to convert an analogue signal into a digital signal.

Step 3: apply wavelet transform by applying $X(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \overline{\psi\left(\frac{t-b}{b}\right)} x(t) dt$

- **Decompose signal into LL, HL, LH, and HH bands by computing the wavelet coefficients as $c_{jk} = [W_{\psi f}](2^{-j}, k2^{-j})$**

Step 4: Input these coefficients to deep learning

- **Process through convolutional layers $n_{out} = \left\lfloor \frac{n_{in} + 2p - k}{s} \right\rfloor + 1$, n_{in} denotes the input attributes, n_{out} denotes the output features, k convolution kernel size, p padding size, s is the stride**
- **Process the convolved data through pooling layer $h_{xy}^l = \max_{i=0, \dots, s, j=0, \dots, s} h_{(x+1)(y+j)}^{l-1}$**
- **Perform linearization by applying linear layer**
- **Apply BiLSTM layer**
- **Process the memory unit data through fully connected layer $z^l = W^l h^{l-1}$**
- **Soft max layer $softmax(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$**

Step 5: obtain the final output speech data and measure the performance

The components of CNN have been discussed in detail in chapter 2. Figure 4.3 depicts an example of the CNN procedure.

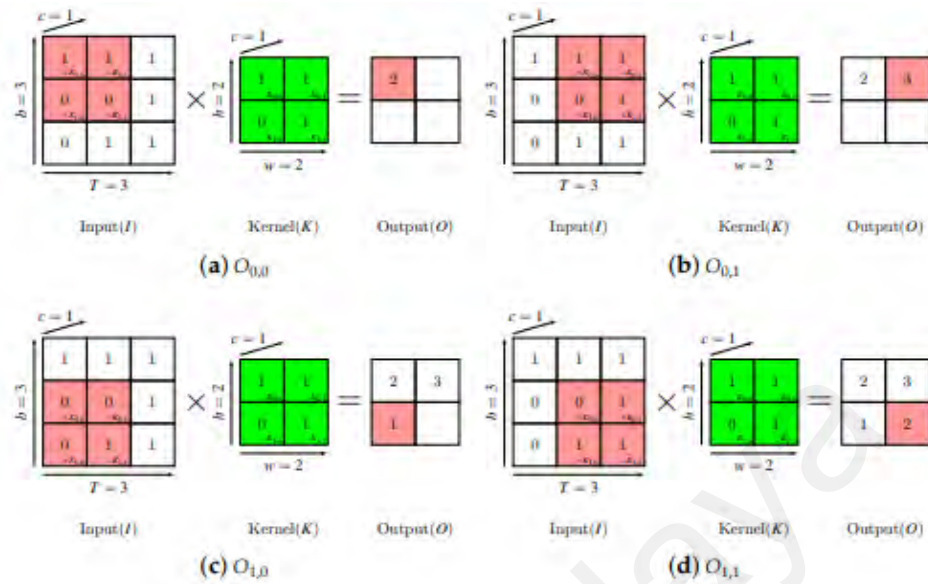


Figure 4.3: An example of CNN process (Dong wang et al., 2019)

a) *LSTM and BiLSTM Layer*

RNNs typically struggle to learn long-term sequences. Hochreiter and Schmidhuber developed LSTM to solve the issue. A memory block in LSTMs, unlike RNNs, consists of a self-hidden unit (memory cell) with a recurrent connection and two gating units (input and output gates) that control access to data in the memory cell based on the previous context. For sequence labelling tasks like offline speech and handwriting recognition, LSTM networks are successfully used. The LSTM can keep or forget information by using these operations:

- *Core Concept*

The cell state and its multiple gates are the basic idea of LSTMs. The cell state works as a highway for the transportation of relative information throughout the entire sequence chain. It can be considered as the network's "memory." The cell state might, in

theory, carry important information when the sequence is processed. Thus, even information from earlier time steps might reach later time steps, decreasing the impact of short-term memory. Information is added to or withdrawn from the cell state via gates as the cell state flows. The gates, which control the information that is allowed on the cell state, are different neural networks. During training, the gates can learn what information is essential to keep or forget.

- *Sigmoid*

Sigmoid activations can be found in gates. The tanh activation is identical to a sigmoid activation. It compresses data between 0 and 1 rather than between -1 and 1. Because any integer multiplied by 0 is 0, values disappear or are "forgotten," making it easy to update or forget data. Any number multiplied by one has the same value, hence that value is "stored" or remains the same. The network can learn which data is important to keep and which should be deleted based on importance.

- *Forget gate*

The forget gate comes first, this gate decides what data should be deleted or kept. The sigmoid function processes data from the previous hidden state as well as data from the current input. There are values between 0 and 1. To forget means closer to 0, and to keep means closer to 1.

- *Input Gate*

Just get the input gate to update the cell state. First, a sigmoid function is used to process the current input and the previous hidden state. By converting the values to be

between 0 and 1, it decides which values will be changed. 1 denotes important, 0 denotes that it is not important. Additionally, the hidden state and current input are passed to the tanh function, which squashes values between -1 and 1 to help the network regulation. The tanh output is then multiplied by the sigmoid output. The information that should be kept from the tanh output will be decided by the sigmoid output.

- *Cell State*

should be able to figure out the cell state. The forget vector is first multiplied pointwise by the cell state. If multiplied by values close to 0, this could result in the cell state losing values. Then, using a pointwise addition, researchers update the cell state to new values that the neural network considers relevant by taking the output from the input gate. Our new cell state is given by that.

- *Output Gate*

The output gate is the final gate. The next hidden state is decided by the output gate. Keep in mind that the hidden state holds data about previous inputs. Predictions are also made using the hidden state. First, a sigmoid function is used to process the current input and the previous hidden state. The newly modified cell state is then passed to the tanh function. To decide what information the hidden state should carry, simply multiply the tanh output by the sigmoid output. The hidden state is the output. The new hidden and cell states are then carried over to the next time step. Figure 4.4 below illustrates the architecture of the LSTM unit.

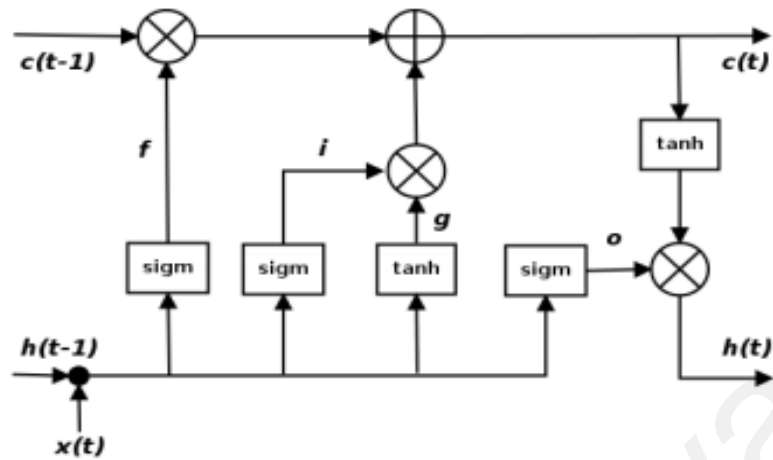


Figure 4.4: LSTM unit (Dong wang et al., 2019)

To improve classification performance, this research further improved the LSTM unit and presented a BiLSTM model. Information flows from backward to forward in unidirectional LSTMs, but bidirectional LSTMs use hidden states to forward information from backward to forward and forward to backward. This enhances how well LSTM network learn. The BiLSTM architecture is illustrated in Figure 4.5.

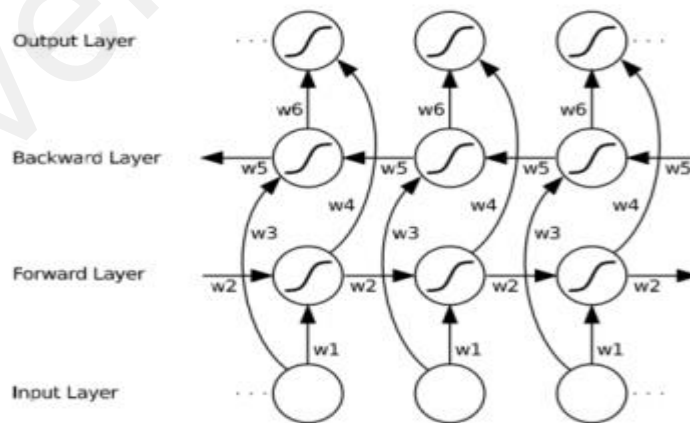


Figure 4.5 BiLSTM architecture (Dong wang et al., 2019)

Table 4.2: Detailed explanation of each layer of CNN with their function, libraries and their work process.

Layer	Layer parameters	Function	Library	Working process
Convolution layer	Kernel size, padding, stride	$n_{out} = \left\lfloor \frac{n_{in} + 2p - k}{s} + 1 \right\rfloor$	convolution2dLayer(filterSize, numFilters, 'Padding', 1)	Convolutional layers work on the input using a convolution operation and send the outcome to the following layer. All the pixels in a convolution's receptive area are converted into a single value.
Batch Normalization	Batch-normalization (BN) is an algorithm which speeds up and stabilises the training of Deep Neural Networks (DNN).	$y^l = f\left(\frac{1}{1-p} \cdot W^l(r^{l-1} * y^{l-1}) + b^l\right)$	layer = batchNormalizationLayer(Name, Value)	The first and second statistical moments (mean and variance) of the current batch are used to normalise activation vectors from hidden layers.
Sigmoid	An activation function is applied to a weighted sum of inputs, and the outcome is used as an input for	$\sigma(x) = 1/(1 + \exp(-x))$	Y = sigmoid(X)	Typically represented by $\sigma(x)$ or sig(x), the sigmoid function is a specific case of the logistic function (x).

	the next layer.			
Pooling	Dimensions of the feature map include height, width, channels, filter size, and stride length.	$S_j = \max_{i \in R_j} a_i$	<code>maxPooling2dLayer(3, 'Stride', 2)</code>	<p>The dimension of the feature maps is reduced by pooling layers. As a result, it reduces the amount of network computation and the number of parameters that must be learned.</p> <p>The feature map generated by a convolution layer's feature pooling layer summarises the features that are present in a certain region.</p>

4.3 Dataset details

There are many kinds of noise that exists other than environment, for example cable noise, attenuation noise, transient noise, white noise, impulse noise, etc. However, this research is focused on environmental noises as speech communication devices are mostly used in outdoor environments. Even though there are many data sources available, this research used AURORA database as it included environmental noises from -5db to 20db SNR range. Thus, it needs low SNR levels of signals as it is mainly focused to improve

recognition performance at -5db, -10db, 0db using proposed framework.

The AURORA database, which is taken from the internationally recognised NOIZEUS database for Multi-Channel Speech Enhancement, was used in this research. This database includes the speech recordings of speakers, three men and three women, reciting 30 sentences from the IEEE sentence database. The University of Texas at Dallas' Speech Processing Lab used Tucker Davis Technology (TDT) to capture each speaker's five words at a sampling frequency of 25 kHz, which was later down sampled to 8 kHz. Each sentence was mixed with different kinds of environmental noises such as Airport, Babble, Car, Exhibition Restaurant and AWGN. To get both clean and noisy signals, this research employed an intermediate reference system (IRS) filters (<https://ecs.utdallas.edu/loizou/speech/noizeus/>). To achieve the appropriate SNR levels, the recovered noise segments were artificially introduced to the clean speech signal.

Similar AURORA dataset has been used for both benchmark and proposed experiments. With respect to the proposal by this research, it required more samples to train CNN layer. Huge datasets are used to train the CNNs, and it is possible that the more data, the more accurate the model would be, otherwise, other processes, like transfer learning, must be used to increase the data. CNN can automatically identify distinctive elements in input signals without actual human involvement. This dataset comprises a total of 30 sentences which are generated by three female and three male participants. These datasets samples are deteriorated by eight types of noise which are generated at varied SNR levels. These noises include airport, restaurant, train, car, babble, and exhibition hall and railway station noise. 30 sentences mixed with each noise at each level of SNR. In total 900 speech noisy

signals were used which was divided into 80% for training and 20% for testing.

4.4 Experimental Design and Setup

4.4.1 Experimental Design

In this segment, the research describes the experimental exploration of the proposed speech enhancement algorithm and compared the obtained performance with existing algorithms.

Experimental design of proposed CNN based Multi-Channel Speech Enhancement is presented in Table 4.3.

Table 4.3: Experimental design of proposed noise filtering framework (CNN based) in Multi-Channel Speech Enhancement

Speech Enhancement	Speech Database	SNR/db	Noises
CNN based Multi-Channel Speech Enhancement	Aurora	-10db, -5db, 0db, 5db, 10db, 15db and 20db	Airport, Babble, Car, Exhibition, restaurant and White gaussian noise

4.4.2 Experimental Setup

4.4.2.1 Sampling Setup

Two transducers output was pre-amplified, fed to a single-stage bandpass filter(80Hz-16KHz), after which the gain was adjusted, and level shifted to 1.75V, and then fed to individual ADC's (analog to digital converter).

This research configures ADCs at 12bit vertical resolution and 16000 Samples per second (+/- 50 due to clock stability). Data written to SD card via Conversion completed the interrupt linked to DMA channel which wrote the value in SD card and a copy in Buffer variable defined in RAM. Both ADC sampling times were synchronized. Amplifiers used were based on LM358 general purpose Opamp.

4.4.2.2 Variability Setup

Timer 1 PWM channels connect the Two 9G servos at 16bit resolution (Effective usable steps were around 30000 per servo due to higher ARM deflection of Servos). The distance between each microphone is fixed at 10 mm.

4.4.2.3 Noise and Sample Utterance System Setup

The primary noise driver is Edifier 2.0 channel speaker. The speech is varied at the amplifier and the noise samples were continuously looped and fed to the amplifier from the BeagleBone Black Single Board.

The speech samples were driven with only left channel of Logitech USB speakers and the BeagleBone Black single-board computer fed the samples.

4.4.2.4 SNR Setup

The desired SNR (-10dB, -5dB, 0dB, 5dB, 10dB, 15dB and 20dB) was achieved by individually tuning the noise sound amplifier gain control and the sample utterance amplifier gain control by measuring individual Sound Pressure levels (SPL) to calculate the values.

4.5. Software Requirements

The proposed approach was implemented by using MATLAB 2021a. This tool is widely adopted for various signal processing tasks such as image processing, speech processing and ECG signals. In this study, the aforementioned tool was utilised for speech processing tasks. The proposed model used the following toolboxes:

- *Audio toolbox:*

Tools for audio processing, speech analysis, and acoustic measurement are offered by Audio Toolbox. It provides algorithms to evaluate the acoustic signal metrics like loudness and sharpness, processing audio signals with normalization and time stretching, and extracting audio properties like MFCC and pitch. To train machine learning and deep learning models, researchers can import, classify, and enhance audio data sets using Audio Toolbox. For high-level semantic analysis of audio recordings, the pre-trained models offered can be used.

- *Data acquisition toolbox:*

For setting data acquisition devices, reading data into MATLAB and Simulink, and publishing data to DAQ analogue and digital output channels, Data Acquisition Toolbox™ offers apps and functions. The toolbox apps enable interactive configuration of a data acquisition interface and hardware configuration. To automate the data collecting, one can then create identical MATLAB code. Analogue input, analogue output, counter/timer, and digital I/O subsystems of a DAQ device can all be freely controlled using toolbox functions. Data collected from several devices can be synchronised and access device-specific functionalities.

Data analysis is available both in real-time and for later processing. Based on the findings of past investigations, researchers can also automate tests and make iterative adjustments to test configuration.

- *Digital signal processing toolbox*

For designing, simulating, and analysing signal processing systems in MATLAB and Simulink, DSP System Toolbox offers algorithms, apps, and scopes. Real-time DSP systems can be modelled for use in communications, radar, audio, medical devices, Internet of Things, and other applications. Researchers can design and analyse FIR, IIR, multirate, multistage, and adaptive filters with DSP System Toolbox. For system development and verification, signals from variables, data files, and network devices can be streamed. Researchers may dynamically visualise and assess streaming signals using the Time Scope, Spectrum Analyzer, and Logic Analyzer. The toolbox provides C/C++ code generation for desktop prototype and deployment to embedded processors,

including ARM, Cortex architectures. Additionally, it supports the production of HDL code from filters, FFT, IFFT, and other algorithms as well as bit-accurate fixed-point modelling.

- *Wavelet toolbox*

For the analysis and synthesis of signals and images, Wavelet Toolbox offers functions and applications. The toolkit includes algorithms for data-adaptive time-frequency analysis, continuous wavelet analysis, wavelet coherence, and synchrosqueezing. Additionally, the toolbox has algorithms and features for wavelet packets and dual tree transforms, as well as for decimated and non-decimated discrete wavelet analysis of signals and images.

Continuous wavelet analysis allows you to analyse the time variation of spectral features, identify common time variation patterns between two signals, and apply time-localized filtering. Researchers can analyse signals and images at various resolutions using discrete wavelet analysis to find changepoints, discontinuities, and other events that aren't immediately visible in raw data. Researchers can do fractal analysis on data to find hidden patterns and compare signal statistics on various scales.

- *Deep learning toolbox*

With algorithms, pre-trained models, and apps, Deep Learning Toolbox offers a platform for developing and implementing deep neural networks into application. Long short-term memory (LSTM) networks and convolutional neural networks (ConvNets, CNNs) can be used to conduct classification and regression on image, time-series, and text data.

Automatic differentiation, unique training loops, and shared weights can be used to create network topologies like generative adversarial networks (GANs) and Siamese networks. Researchers may graphically create, evaluate, and train networks with the Deep Network Designer software. Researchers may manage many deep learning experiments, keep track of training parameters, examine outcomes, and compare code from several experiments using the Experiment Manager tool. Layer activations are visible, and training progress is graphically monitored.

4.6 Performance measurement parameters

The proposed approach's results are quantified using the following metrics: Perceptual Evaluation of Speech Quality (PESQ), Cepstrum Distance Measures (CEP), Mean Opinion Score (MOS), Frequency Weighted SNR, Short Time Objective Ineligibility measure (STOI), Mean of SNR, Means of Segmented SNR, signal distortion (Csig), Cbak, a compound estimate for noise distortion, a compound estimate for overall speech quality (Covl), Mean LLR and Itakura-Satio. The following can be used to compute these parameters:

- PESQ: ITU-T advises using the PESQ measurement, a complex parameter, to evaluate speech quality. The average asymmetrical disturbance A_{ind} and average disturbance D_{ind} are combined linearly to create the PESQ. Researchers can calculate this as:

$$PESQ = a_0 + a_1 D_{ind} + a_2 A_{ind} \quad (4.15)$$

Where a_0 , a_1 and a_2 are the three constant parameters whose values are 4.5, -0.1 and -0.0309

- Log-likelihood ratio (LLR): the LLR is computed as

$$d_{LLR}(\vec{a}_p, \vec{a}_c) = \log \left(\frac{\vec{a}_p R_c \vec{a}_p^T}{\vec{a}_c R_c \vec{a}_c^T} \right) \quad (4.16)$$

Where \vec{a}_c denotes the LPC vector obtained from original speech frame, \vec{a}_p denotes the LPC vector obtained from the enhanced speech frame and R_c represents the autocorrelation of original speech signal.

The algorithm of linear predictive coding (LPC), which makes use of the data from a linear predictive model, is mostly employed in audio signal processing and speech processing to capture the spectral range of a digital signal of speech in compressed form. For example, the below figure shows LPC vector calculation in original signal in a matlab simulator.

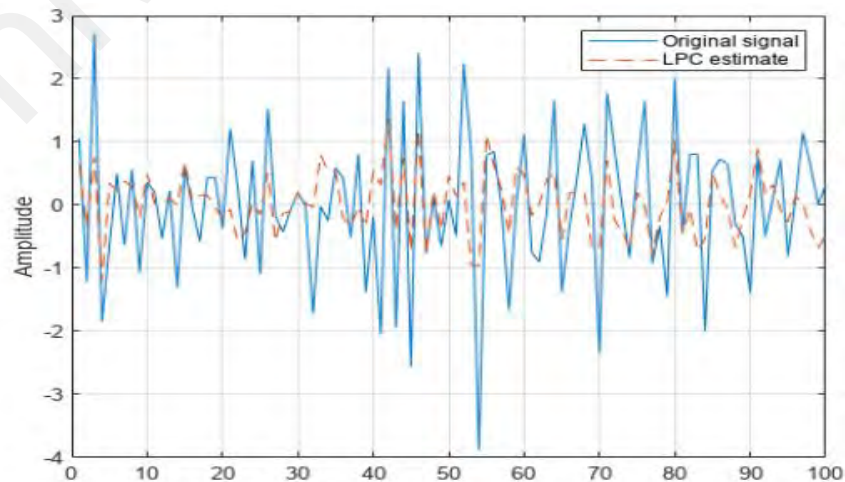


Figure: LPC vector estimation in original speech signal in matlab simulator

- Itakura-Saito (IS): the IS parameter can be computed as follows:

$$d_{IS}(\vec{a}_p, \vec{a}_c) = \frac{\sigma_c^2}{\sigma_p^2} \left(\frac{\vec{a}_p R_c \vec{a}_p^T}{\vec{a}_c R_c \vec{a}_c^T} \right) + \log \left(\frac{\sigma_c^2}{\sigma_p^2} \right) - 1 \quad (4.17)$$

Where σ_c is the LPC gain of clean signal whereas and σ_p represents the LPC gains of enhanced speech signal.

- Cepstrum coefficients: the CC can be obtained as follows:

$$d_{CEP}(\vec{c}_c, \vec{c}_p) = \frac{10}{10 \log 10} \sqrt{2 \sum_{k=1}^p [c_c(k) - c_p(k)]^2} \quad (4.18)$$

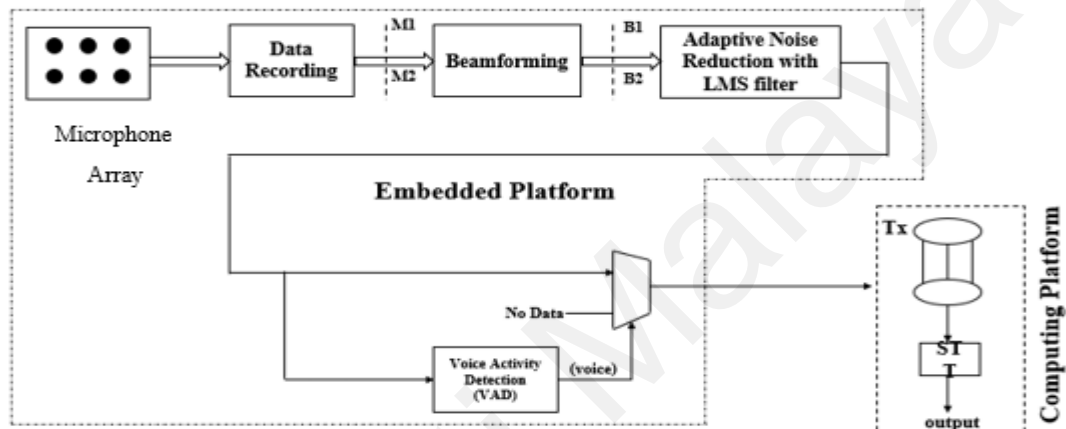
Where \vec{c}_c denotes the LPC gain of clean signal and \vec{c}_p represents the LPC gains of enhanced speech signal.

- MOS (mean opinion score) MOS assigns a value to the overall quality of the delivered speech through a network in comparison with the original speech. MOS ratings have a range from 1 (bad) to 5 (excellent) (Ramana A V et al., 2012).
- STOI (Short time objective intelligibility): Intelligibility measure which is highly compared with intelligibility of degraded speech signals. STOI ratings have a range from -0.5 to 4.5 (Higher the value implies better quality) (Taal H C et al., 2011).
- ***Speech Signal distortion, Background noise distortion and Overall Quality***

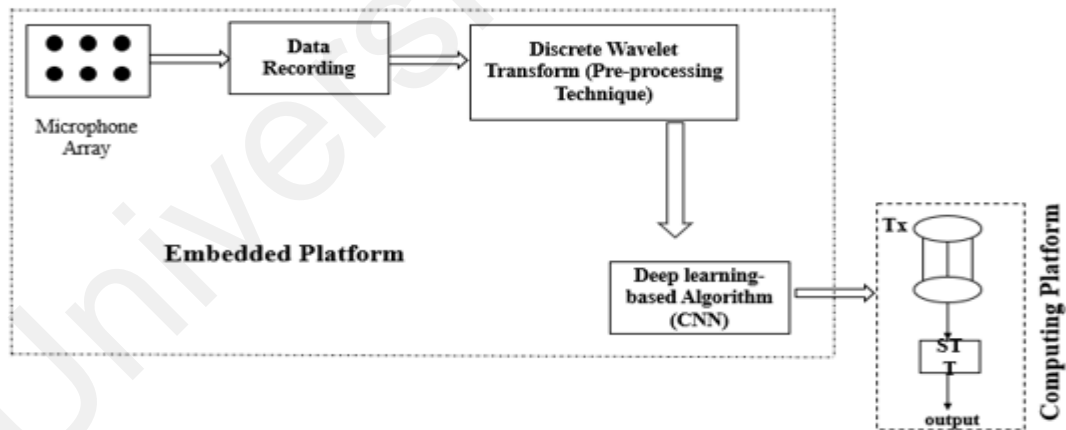
The assessments should be based on the overall quality on the speech signal, the environmental noise, or both. This algorithm tells the listener to pay attention to and rate the improved speech signal in turn (Hu and Loizou 2006).

4.7 The differences between existing MCSE and the proposed deep learning-based noise filtering framework.

Figure 4.6 illustrates the difference between existing Multi-Channel Speech Enhancement system and proposed deep learning based noise filtering framework in MCSE.



4.6 (a) Existing Multi-Channel Speech Enhancement (MCSE) (Alessandro et al., 2017)



4.6 (b) Proposed Multi-Channel Speech Enhancement (MCSE) (Alessandro et al., 2017)

Figure 4.6 Existing and Proposed Multi-Channel Speech Enhancement Framework

The existing MCSE consists of wearable microphones with beamforming, adaptive noise reduction and voice activity detection algorithms. The existing researches have

focused on improving the performance of these speech communication devices. Several algorithms have been presented to improve the speech quality in MEMS microphones, but these algorithms suffer from low performance. To overcome this problem, this research proposed noise filtering framework using DWT preprocessing algorithm and deep learning-based CNN-BLSTM algorithms.

4.8 Summary

This chapter presented a proposed solution for Multi-Channel Speech Enhancement by using wavelet preprocessing algorithm and deep learning algorithm. In the first phase, the research captured the noisy speech signals from the speech communication device. After capturing the speech signals, the speech signals were stored to process for further enhancement operation. In the next phase, this signal was passed through the wavelet transform model where the signal is decomposed into multiple wavelet bands such as LH, HL, HH and LL bands. These bands were processed through the wavelet coefficient computation to obtain the attributes. Later, obtained attributes were fed into the deep learning architecture where the noisy signals were processed through multiple deep learning layers to obtain the final output. Based on this proposed approach, this research presented the evaluation of proposed approach in the next chapter where various results and outcomes of this work are described.

CHAPTER 5: EVALUATION, RESULTS AND DISCUSSION

5.1 Overview

This chapter presents the results and discussion on the evaluation conducted using the proposed framework. This research measures the performance of the proposed algorithms used in the framework in terms of various parameters such as SNR, log likelihood ratio, perceptual evaluation of speech quality, cepstrum distance measures, mean opinion score, frequency weighted SNR, short term objective intelligibility and Itakura –Saito (IS). Moreover, this research also measured the performance of the proposed Multi-Channel Speech Enhancement system as regards word recognition rate and word error rate to show the robustness of this method for varied type of noise considered for speech communication devices.

5.2 Evaluation

In this work, the spectrogram analysis and the Word Recognition Rate (WRR) were used to evaluate the performance of the Multi-Channel Speech Enhancement system in a noisy environment (stationary and non-stationary noise).

5.2.1 Spectrogram Analysis

Spectrogram analysis is used to analyse the amplitude of speech signals (S Haykin et al., 1991). The spectrogram analysis for both stationary (White Gaussian Noise) and non-stationary environmental noises (Babble, Airport, Car, Exhibition, and Restaurant) were performed on time-domain using MATLAB.

5.2.2 Word Error Rate (WER)

This research tested the voiced speech signal received after the voice activity detection with the ASR speech to text engine to determine the word error rate (WER). Word error rate is calculated to evaluate the performance of the Multi-Channel Speech Enhancement systems. WER is computed as follows:

$$WER = \frac{S+D+I}{N} \quad (5.1)$$

Where N is the total number of words/letters in the sentence, S is the number of substitutions of other words, D is the number of deletions and I is the number of insertions in a sentence.

By calculating the WER, the word recognition rate (WRR) is determined as:

$$WRR = 1 - WER \quad (5.2)$$

WRR measures the performance accuracy of Multi-Channel Speech Enhancement system.

5.2.3 Performance measurement parameters

In this research, some of the metrics were also used to measure the results of the proposed approach, which included PESQ (Perceptual Evaluation of Speech Quality), Mean Opinion Score (MOS), CEP (Cepstrum Distant Measures), Frequency Weighted SNR, Short Time

Objective Ineligibility (STOI), Mean of Segmented SNR (SNRseg), Signal Distortion (Csig), Cbak, a composite assessment for background noise distortion (Cbak), a composite assessment for overall speech quality (Covrl), Mean Log-likelihood Ratio (LLR), and Itakura-Saito (IS). Table 5.1 summarizes the results of the test in more detail.

Table 5.1: Description of the parameters for performance evaluation

Parameter	Measuring quantity	Range	Description
PESQ	Speech Quality	-0.5 - 4.5	Higher the value implies better quality
MOS	Speech Quality	1-5	Higher is better.
CEP	Error	[0-10]	Minimum is better
Frequency weighted SNR	Quality of Speech	typically, between 10 and 35 dB on average	Higher is better.
STOI	Speech Quality	-0.5 - 4.5	Higher the value implies better quality
SIG	Distortion	1-5	Higher is better.
BAK	Distortion	1-5	Higher is better.
OVRL	Speech quality	1-5	Higher is better.
LLR	Speech quality	1-5	Higher is better.
IS	Distance	1-5	Higher is better.

5.3 Results

5.3.1 Spectrogram Analysis for Multi-Channel Speech Enhancement

A spectrogram analysis shows the spectral illustrations of a time-varying signal (Flanagan et al., 1972). Figure 5.1 represents the clean speech signal, while Table 5.2 presents the spectrograms analysis: (a) noisy speech at different levels of SNR and (b) enhanced speech using proposed Deep learning -based Multi-Channel Speech Enhancement.

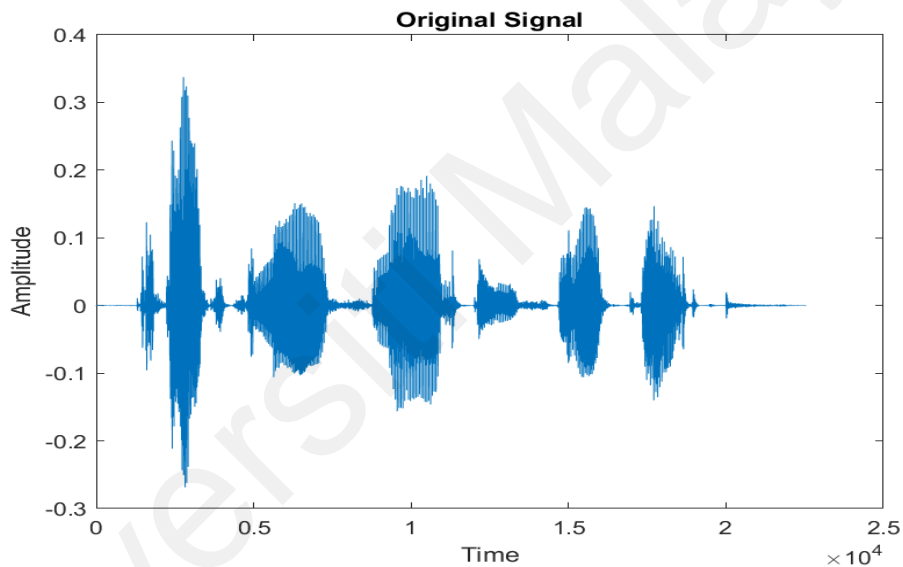
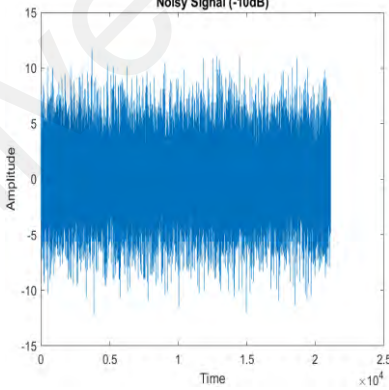
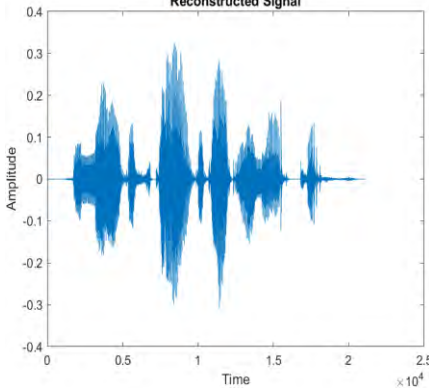


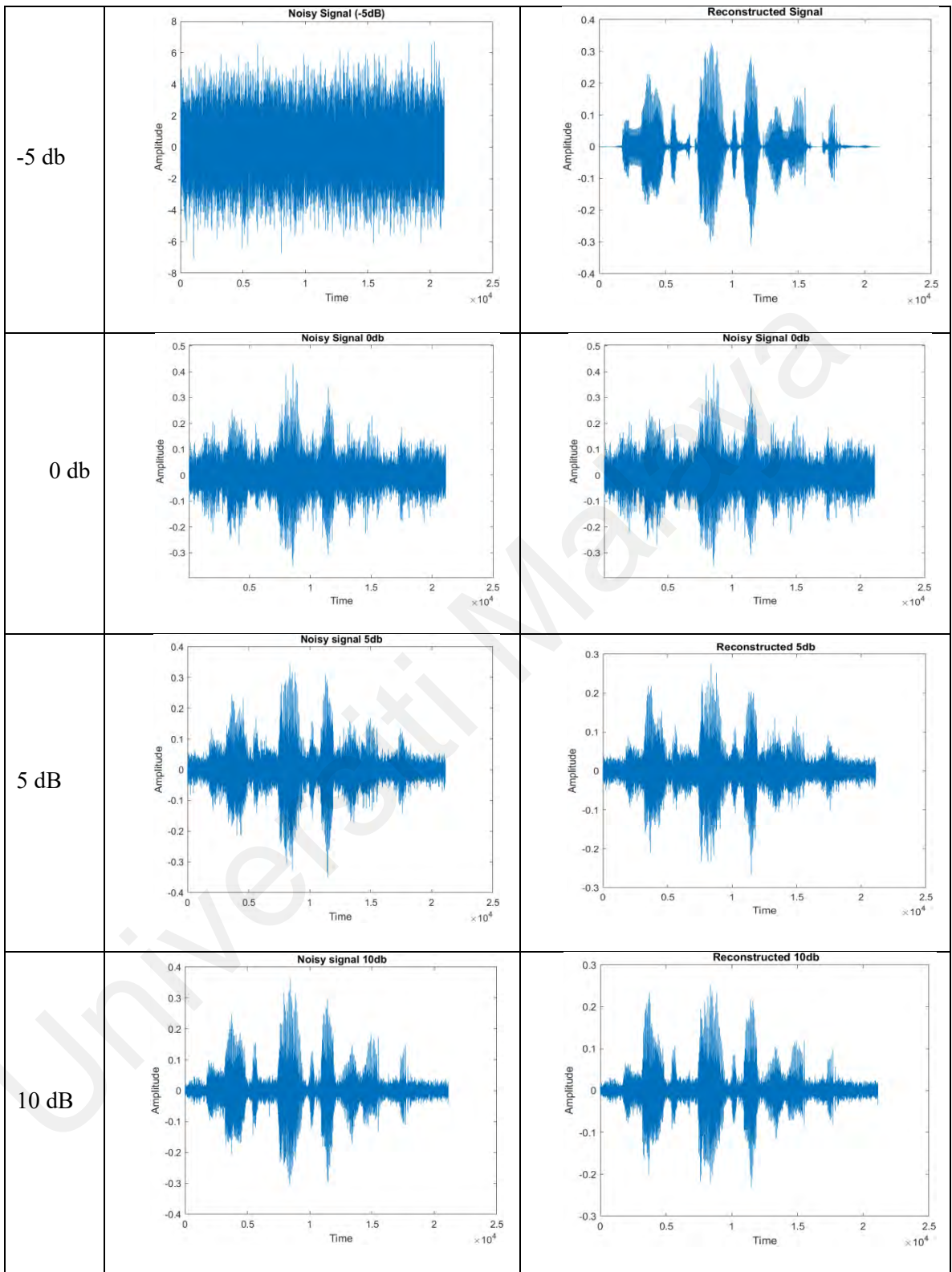
Figure 5.1: Clean speech signal

This research considered airport noise, babble noise, restaurant and white gaussian noise for this analysis. The experiment added the -10db, -5db, 0db, 5db, 10db and 15db noise to the original signal and processed it through the considered speech enhancement system.

Table 5.2 presents the unfiltered and the reconstruction quality of the proposed deep learning based MCSE at -10db, -5db,0db, 5db,10db and 15db noise for the airport noise. Table 5.3 presents the unfiltered and the reconstruction quality of the proposed Deep learning-based MCSE at -10db, -5db,0db, 5db,10db and 15db noise for the babble noise. Table 5.4 presents the unfiltered and the reconstruction quality of the proposed Deep learning based MCSE at -10db, -5db, 0db, 5db,10db and 15db noise for the restaurant noise. Table 5.5 presents the unfiltered and the reconstruction quality of the proposed Deep learning-based MCSE at -10db, -5db, 0db, 5db,10db and 15db noise for the AWGN noise which is non-stationary.

Table 5.2: Spectrogram analysis of Airport noisy and enhanced speech signal at different SNR's using proposed Deep learning approach

SNR	Noisy Speech Signal	Enhanced Signal/ Reconstructed Sinal
-10 db		



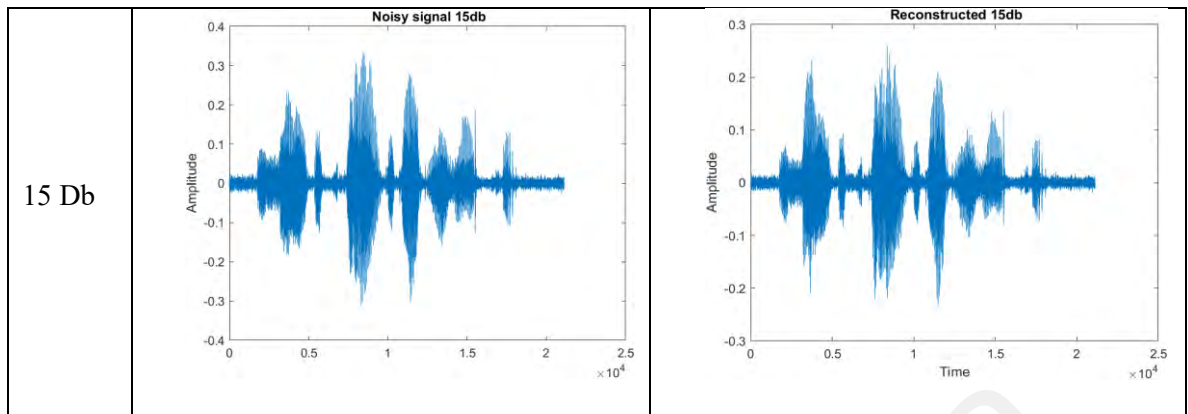
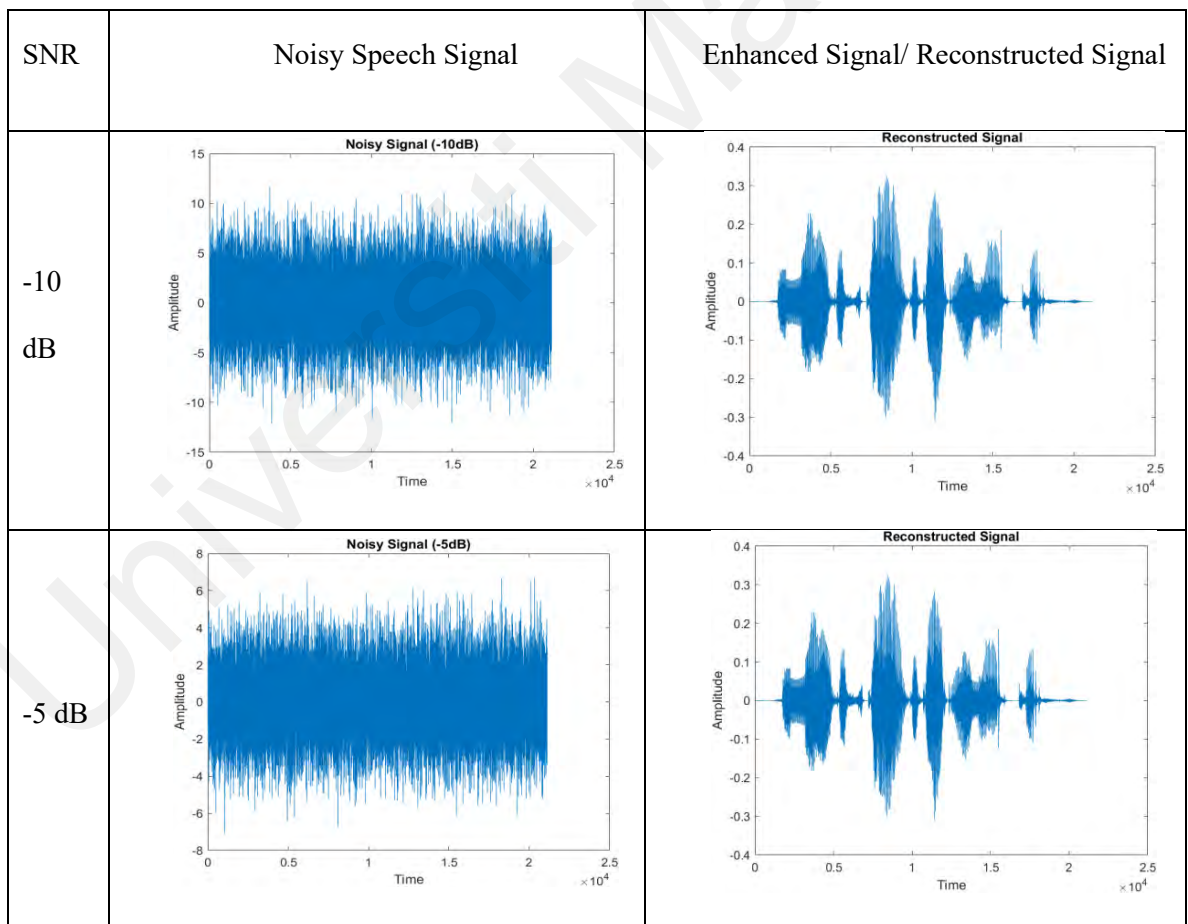


Table 5.3: Spectrogram analysis of Babble noisy and enhanced speech signal at different SNR's using proposed deep learning algorithms



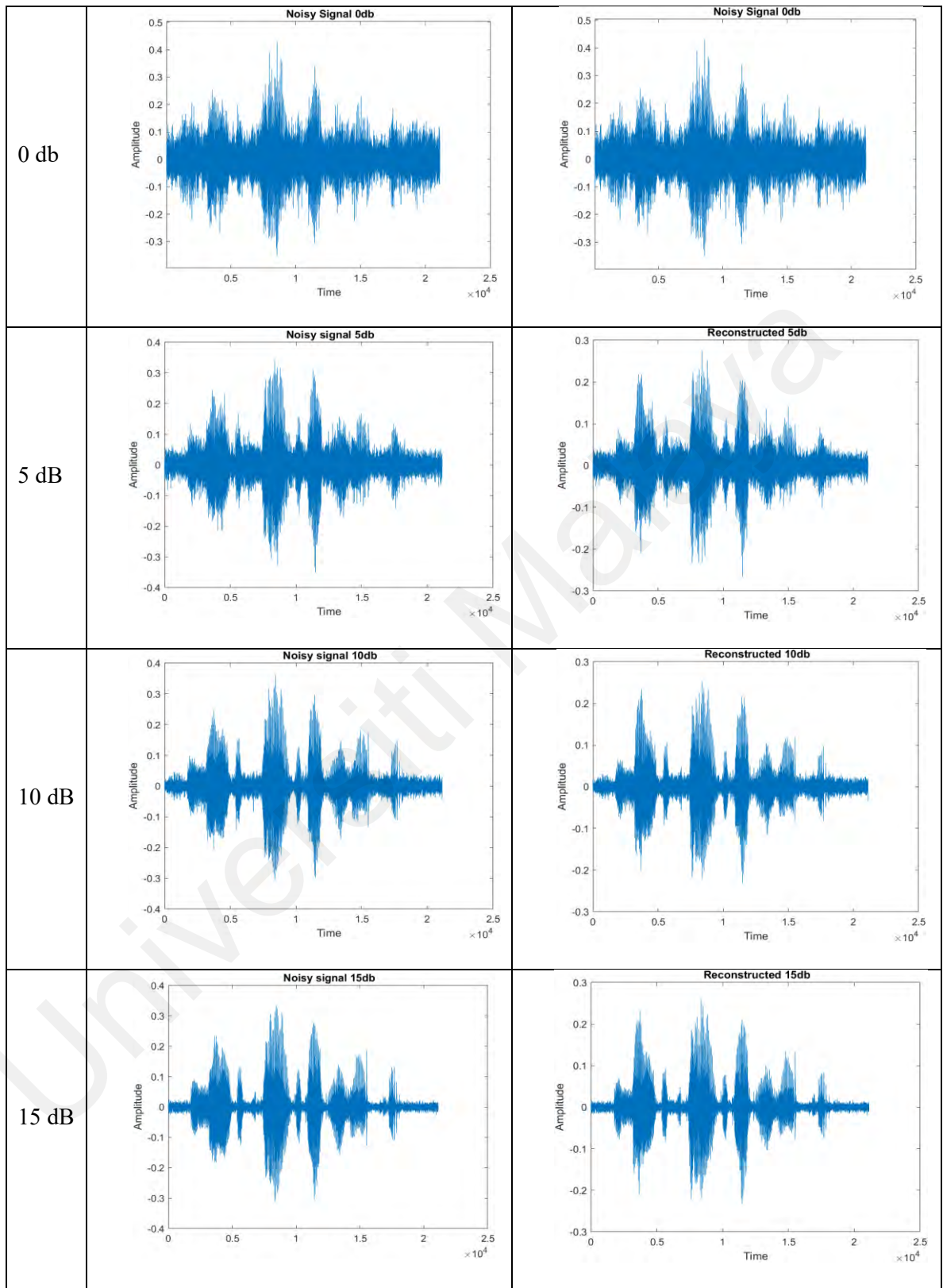
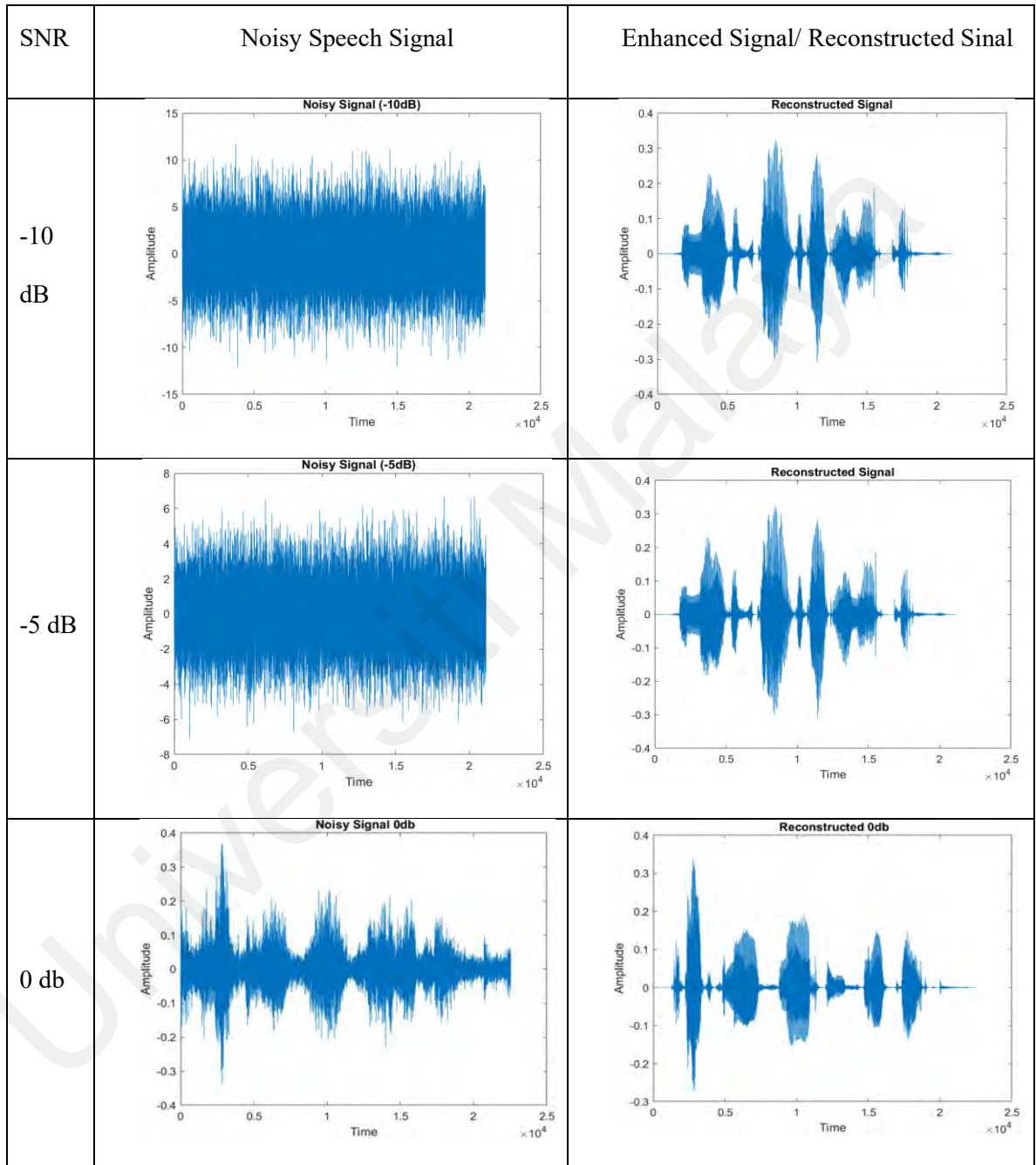


Table 5.4: Spectrogram analysis of Restaurant noisy and enhanced speech signal at different SNR's using the proposed deep learning algorithms.



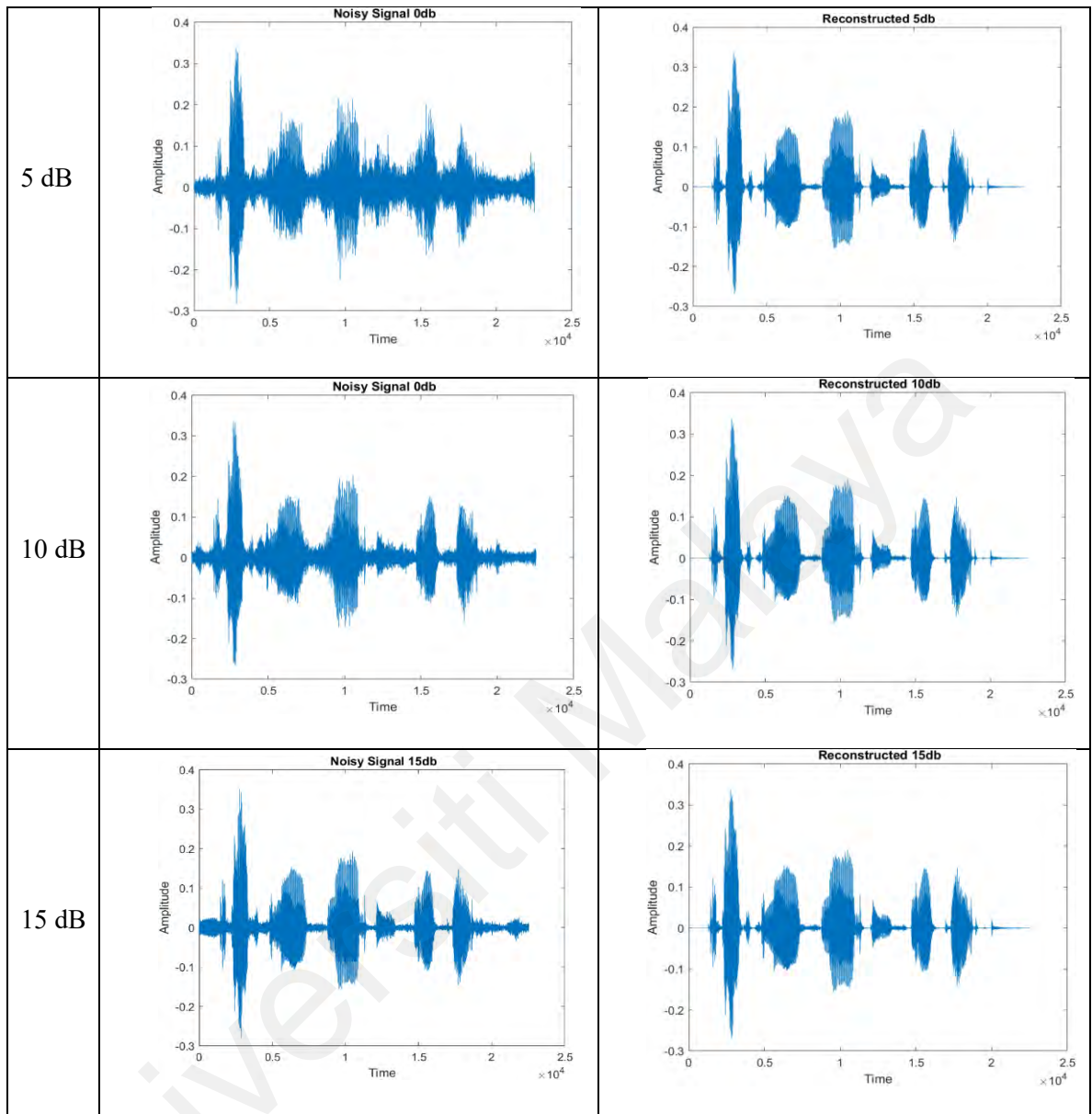
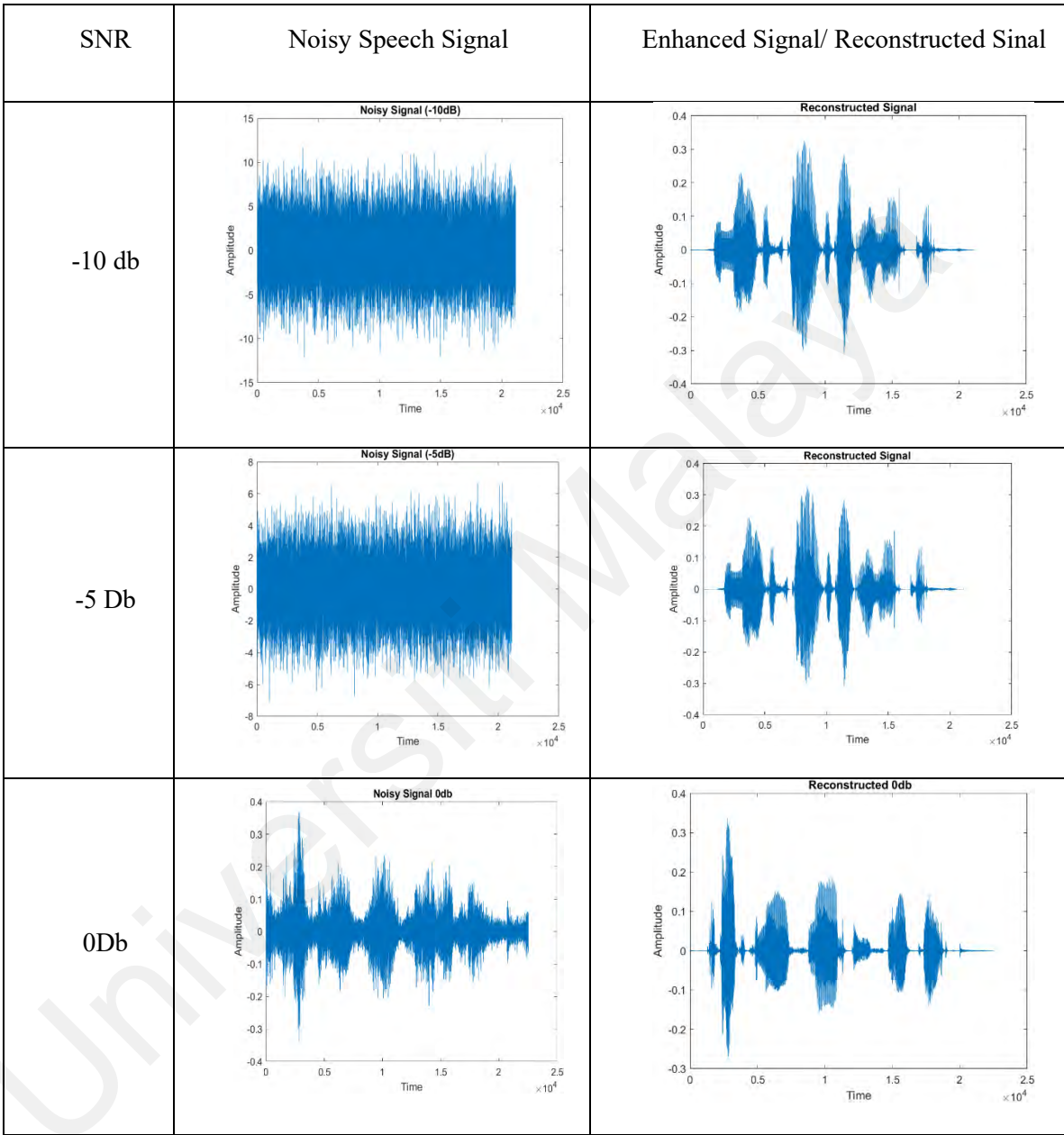
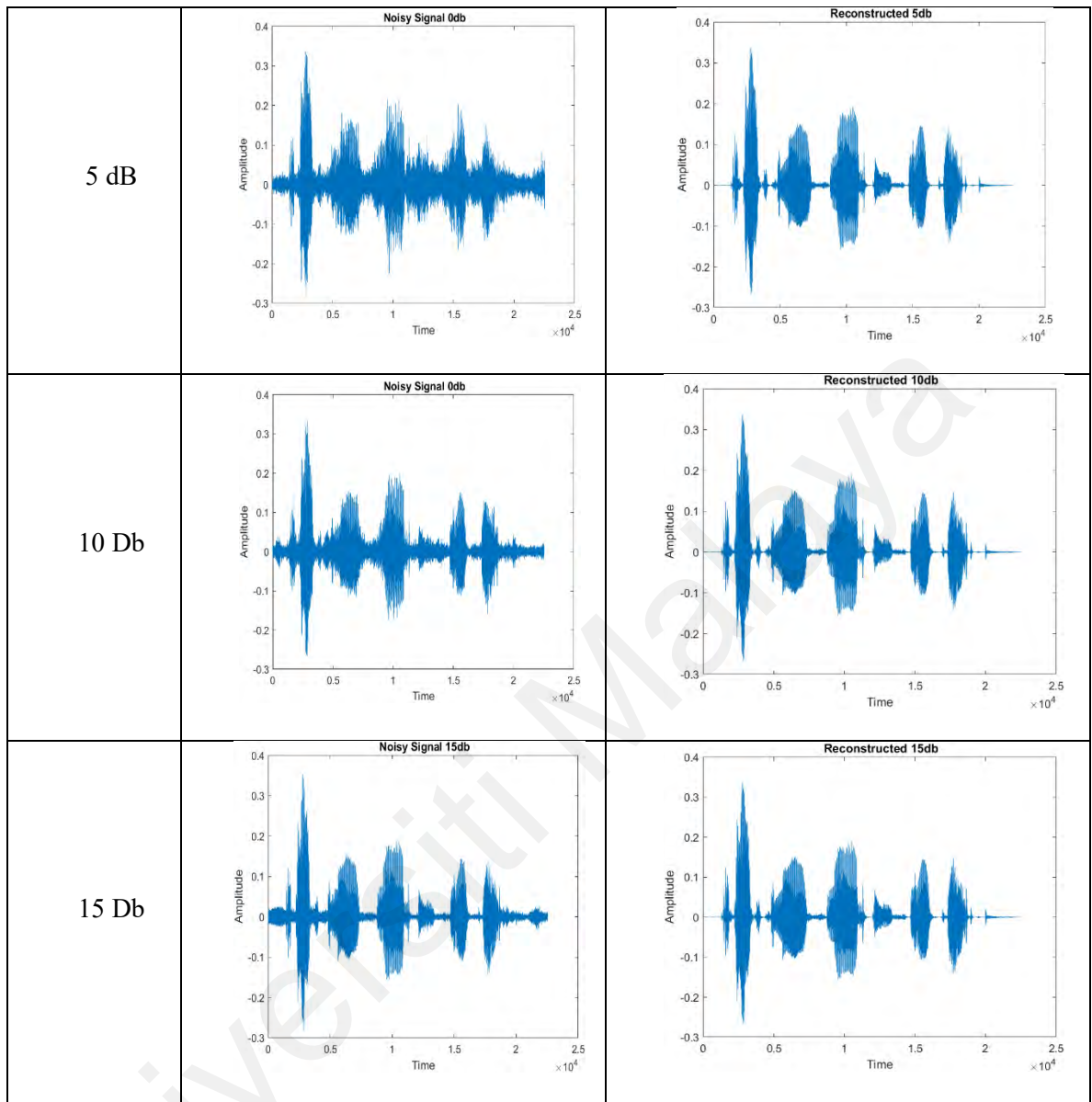


Table 5.5: Spectrogram analysis of Additive white gaussian noise (AWGN-non stationary) noisy and enhanced speech signal at different SNR's using the proposed Deep learning approach





5.3.2 Word recognition rate (WRR) for the proposed Deep learning-based Multi-Channel Speech Enhancement system

Table 5.6 presents the evaluation outcomes of the experiments using the proposed Deep learning based Multi-Channel Speech Enhancement at different levels of SNRs under stationary and non-stationary noisy environments.

Table 5.6: WRR and WER performance by using proposed Deep learning approach

Noise	SNR	WRR	WER
Airport	-10 dB	70.55	26.82
	-5 dB	72.51	25.20
	0 dB	78.75	21.25
	5 dB	77.44	22.56
	10 dB	67.15	32.85
	15 dB	75.44	24.56
Babble	-10 dB	68.50	32.25
	-5 dB	70.25	31.20
	0 dB	70.32	29.68
	5 dB	66.44	33.56
	10 dB	73.79	26.21
	15 dB	61.49	38.51
Car	-10 dB	72.50	26.50
	-5 dB	74.60	25.20
	0 dB	80.49	19.51
	5 dB	77.44	22.56
	10 dB	81.49	18.51
	15 dB	78.8	21.2
Exhibition	-10 dB	65.15	32.25

	-5 dB	68.25	30.20
	0 dB	80.45	19.55
	5 dB	77.73	22.27
	10 dB	76.42	23.58
	15 dB	73.75	26.25
Restaurant	-10 dB	72.50	26.80
	-5 dB	73.50	25.50
	0 dB	80.39	19.61
	5 dB	77.75	22.25
	10 dB	76.35	23.65
	15 dB	74.48	25.52
AWGN	-10 dB	72.51	27.58
	-5 dB	73.55	26.52
	0 dB	79.78	20.22
	5 dB	75.79	24.21
	10 dB	74.38	25.62
	15 dB	73.48	26.52

Table 5.7: Average results of WRR performance on both stationary and non-stationary by using proposed Deep learning approach

WRR	Stationary	Non-Stationary
-10 dB	69.84	72.51
-5 dB	71.822	73.55

0 dB	78.08	79.78
5 dB	75.36	75.79
10 dB	75.04	74.38
15 dB	72.792	73.48

5.3.3. Performance measurement parameters

Table 5.8 to 5.11 presents the performance analysis for Airport noise, Babble noise, Restaurant noise and AWGN noise using the proposed Deep learning based MCSE respectively.

Table 5.8: Performance analysis for airport noise using proposed Deep learning based MCSE

Noise Type	Parameters	Noise Level					
		-10 dB	-5 dB	0 dB	5 dB	10 dB	15 Db
Airport	PESQ	1.65	1.98	1.85	2.31	2.95	3.85
	MOS	2.19	2.51	2.85	3.1	3.45	4.1
	CEP	8.46	8.82	8.9	9.2	9.65	9.8
	SNRseg	24.25	24.68	26.55	28.59	29.35	32.26
	STOI	0.23	0.26	0.28	0.31	0.39	0.42
	Csig	2.6	2.68	2.95	3.22	3.85	4.23
	Cbak	2.5	2.68	2.65	3.12	3.95	4.55
	Covl	1.42	1.85	3.1	3.6	4.2	4.8
	LLR	0.155	0.218	0.85	0.89	0.87	0.95
	IS	22.55	26.15	26.39	32.25	35.56	36.58

Table 5.9: Performance analysis for Babble noise using proposed Deep learning-based MCSE

Noise Type	Parameters	Noise Level					
		-10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB
Babble	PESQ	2.4	2.46	2.55	3.12	3.68	4.33
	MOS	2.8	3.1	3.2	3.56	4.01	4.61
	CEP	7.6	8.3	8.9	9.3	9.5	9.8
	SNRseg	28.55	29.15	29.35	32.25	35.39	36.55
	STOI	0.18	0.23	0.35	0.42	0.46	0.46
	Csig	2.10	2.46	3.25	3.85	4.2	4.39
	Cbak	2.15	2.35	3.2	3.95	4.3	4.4
	Covl	1.40	1.35	3.25	3.46	3.89	4.6
	LLR	0.15	0.198	0.89	0.95	0.96	0.96
	IS	18.50	21.20	31.25	33.50	36.9	39.56

Table 5.10: Performance analysis for Restaurant noise using proposed Deep learning-based MCSE

Noise Type	Parameters	Noise Level					
		-10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB
Restaurant	PESQ	1.75	1.95	3.1	3.32	4.2	4.65
	MOS	2.12	2.16	3.35	4.21	4.35	4.56
	CEP	8.12	6.22	8.85	9.2	9.45	9.66
	SNRseg	22	21	31.25	33.56	34.58	38.51
	STOI	0.21	0.233	0.36	0.42	0.44	0.49

	Csig	2	2	3.85	4.12	4.25	4.65
	Cbak	2	2	3.3	3.86	4.3	4.4
	Covl	1.51	2.21	3.45	3.89	4.3	4.5
	LLR	0.158	0.319	0.91	0.95	0.94	0.96
	IS	20.25	26.50	33.25	36.25	34.89	39.55

Table 5.11: Performance analysis for AWGN noise using proposed Deep learning-based MCSE

Noise Type	Parameters	Noise Level					
		-10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB
AWGN	PESQ	2.25	2.95	3.25	3.48	4.36	4.90
	MOS	2.85	3.10	3.65	4.12	4.55	4.82
	CEP	7.5	8.1	8.9	9.15	9.35	9.4
	SNRseg	30.25	31.20	32.65	35.25	36.29	38.51
	STOI	0.31	0.34	0.39	0.43	0.44	0.48
	Csig	3.25	3.56	3.91	4.16	4.33	4.85
	Cbak	3.15	3.25	3.39	3.95	4.41	4.56
	Covl	3.3	3.65	4.1	4.25	4.39	4.55
	LLR	0.85	0.86	0.92	0.94	0.96	0.96
	IS	29.55	31.29	34.55	36.51	38.51	39.10

5.4. Performance comparison with existing methods

Table 5.12 and 5.13 presents the performance comparison between the proposed MCSE with existing MCSE system Under Stationary noises at different levels of SNR in terms of WRR.

Table 5.12: Comparison of proposed MCSE with Existing MCSE system Under Stationary noises at different levels of SNR

	SNR/dB	Existing MCSE (beamforming+ANR+VAD)	Proposed MCSE (DWT + CNN- BLSTM)
		WRR	WRR
Stationary Noise	-10dB	42.6	72.51
	-5dB	58.3	73.55
White Gaussian noise	0dB	63.4	79.78
	5dB	68.5	75.79
	10dB	72.6	74.38
	15dB	74.8	73.48

Table 5.13: Comparison of proposed MCSE with Existing MCSE system Under Stationary noises at different levels of SNR

Noise in dB	Existing MCSE (beamforming+ANR+VAD)					Proposed MCSE (DWT + CNN-BLSTM)				
	Non- stationary Noises					Non-stationary Noises				
SNR levels	Airport	Babble	Car	Exhibition	Restaurant	Airport	Babble	Car	Exhibition	Restaurant
-10	5.82	4.04	7.26	6.54	4.54	70.55	68.50	72.50	65.15	72.50
-5	12.32	7.12	13.26	11.54	6.38	72.51	70.25	74.60	68.25	73.50
0	19.06	17.56	16.55	20.23	12.12	78.75	70.32	80.49	80.45	80.39
5	36.14	35	35.16	44.66	38.24	77.44	66.44	77.44	77.73	77.75
10	67.26	74.18	67.2	77.72	55.56	67.15	73.79	81.49	76.42	76.35
15	88.88	90.64	92.02	91.46	75.2	75.44	61.49	78.8	73.75	74.48

Universiti Malaya

5.4 Discussion

This research found out that the real problem existed in the MCSE system from the benchmark experiment. It suggests that the existing methods for noise reduction in MCSE, such as fixed beamforming, ANR, and VAD, are not adequate.

From the proposed experiment and evaluation, Table 5.2 presents the unfiltered and reconstruction quality of the proposed Deep learning-based MCSE at -10db, -5db, 0db, 5db, 10db and 15db noise for the airport noise. Table 5.3 presents the unfiltered and the reconstruction quality of the proposed Deep learning-based MCSE at -10db, -5db, 0db, 5db, 10db and 15db noise for the babble noise. Table 5.4 presents the unfiltered and reconstruction quality of the proposed Deep learning based MCSE at -10db, -5db, 0db, 5db, 10db and 15db noise for the restaurant noise. Table 5.5 presents the unfiltered and reconstruction quality of the proposed Deep learning based MCSE at -10db, -5db, 0db, 5db, 10db and 15db noise for the AWGN noise.

The proposed noise filtering framework achieved the results of 70.55% of WRR at -10db and 78.75% at 0db for airport noise, 68.55% of WRR at -10db and 73.79% at 10db for babble noise, 72.50% of WRR at -10db and 81.49% at 10db for car noise, 65.15% of WRR at -10db and 80.45% at 0db for exhibition noise, 72.50% of WRR at -10db and 75.75% at 0db for restaurant noise and stationary noise which is gaussian noise, resulted to 72.51% at -10db and 79.78 at 0db SNR.

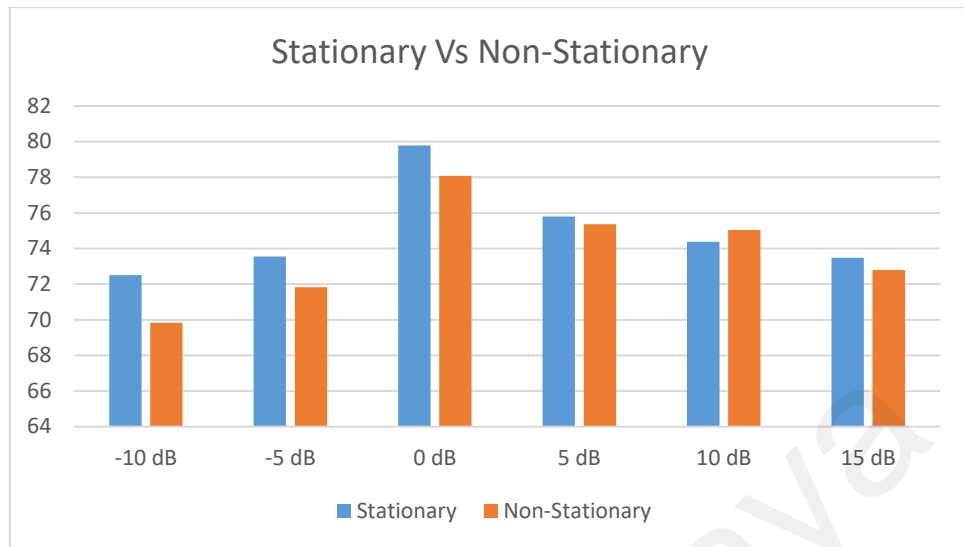


Figure 5.2 WRR for both stationary and non-stationary noises

Figure 5.2 depicts the differences in the WRR for both the stationary and non-stationary noises for proposed MCSE based on DWT preprocessing and CNN-BLSTM. The selected methods work well for both stationary and non-stationary at -10dB to 15dB. For all the dB's, the proposed MCSE is very effective for recognizing speech in both noisy environments. Finally, to determine whether the result for stationary and non-stationary noise was significantly different, the researcher conducted the Analysis of Variance (ANOVA), and the result is shown in Figure 5.3.

ANOVA						
Source of Variation	SS	D f	MS	F	P -value	F crit
Between Groups	1.5101	1	1.5101	0.23525	0.64065	5.31765
Within Groups	51.351	8	6.41896	6	6	5
Total	52.861	9				

Figure 5.3 Results of ANOVA.

From Figure 5.3, the result is not significant, which means that the noise reduction works the same for both stationary and non-stationary.

The performance measurement parameters on the proposed noise filtering framework produces results with the highest PESQ value being 3.85 MOS is 4.1, CEP is 9.8, SNRseg is 32.26, STOI is 0.42, Csig is 4.23, Cbak is 4.55, Covl is 4.8, LLR is 0.95 and IS is 36.58 at 15db SNR and lowest PESQ value is 1.65, MOS is 2.19, CEP is 8.46, SNRseg is 24.25, STOI is 0.23, Csig is 2.6, Cbak is 2.5, Covl is 1.42, LLR is 0.155 and finally IS resulted to 22.55 at -10db in airport noise.

In Babble noise, highest PESQ value is 4.3 MOS is 4.61, CEP is 9.8, SNRseg is 36.55, STOI is 0.46, Csig is 4.39, Cbak is 4.4, Covl is 4.6, LLR is 0.96 and IS is 39.56 at 15db SNR and lowest PESQ value is 2.4, MOS is 2.8, CEP is 7.6, SNRseg is 28.55, STOI is 0.18, Csig is 2.10, Cbak is 2.15, Covl is 1.40, LLR is 0.15 and finally IS resulted to 18.50 at -10db.

In Restaurant noise, highest PESQ value amounted to 4.65, MOS is 4.56, CEP is 9.66, SNRseg is 38.51, STOI is 0.49, Csig is 4.65, Cbak is 4.4, Covl is 4.5, LLR is 0.96 and IS is 39.55 at 15db SNR and lowest PESQ value is 1.75, MOS is 2.2, CEP is 8.12, SNRseg is 22, STOI is 0.21, Csig is 2, Cbak is 2, Covl is 1.51, LLR is 0.158 and finally IS resulted to 20.25 at -10db.

In White Gaussian noise, highest PESQ value is 4.90 MOS is 4.82, CEP is 9.4, SNRseg is 38.51, STOI is 0.48, Csig is 4.85, Cbak is 4.56, Covl is 4.55, LLR is 0.96 and finally IS is 39.10 at 15db SNR and lowest PESQ value is 2.25, MOS is 2.85, CEP is 7.5, SNRseg is

30.25, STOI is 0.31, Csig is 3.25, Cbak is 3.15, Covl is 3.3 , LLR is 0.85 and finally IS is 29.55 at -10db.

By comparing the performance of the developed noise filtering framework in filtering various SNR of environmental noises, findings revealed a WRR of 70.55% at -10dB SNR and 75.44 % at 15dB SNR, while 5.82 % at -10dB and 88.8% at 15dB by the existing MCSE system.

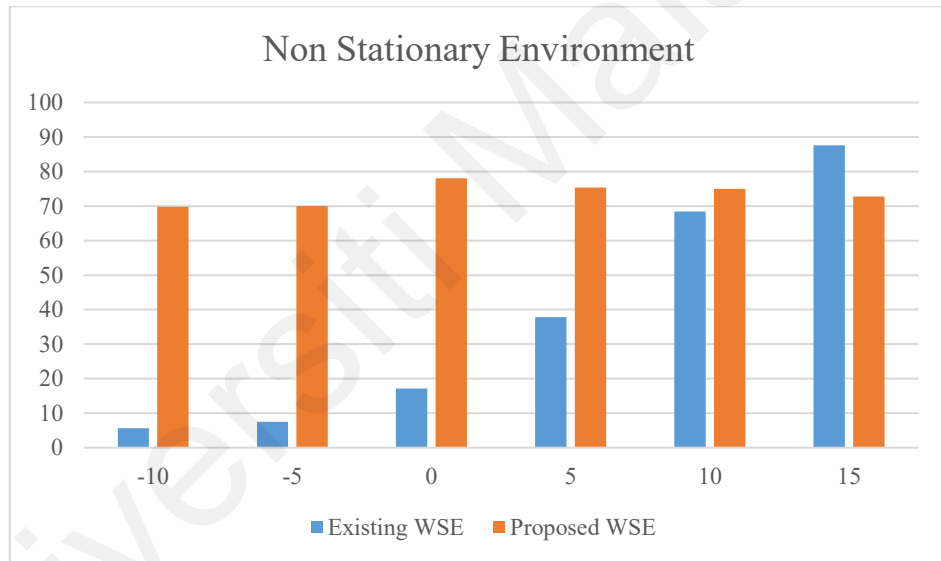


Figure 5.4: WRR of Existing Vs Proposed MCSE under nonstationary environment

Figure 5.4 depicts the differences in the WRR for both the existing MCSE and proposed MCSE under non-stationary noises. The proposed MCSE is very effective for recognizing speech in non-stationary noisy environments compare to existing MCSE. Finally, to determine whether the result for existing MCSE and proposed MCSE was significantly

different, the researcher conducted the Analysis of Variance (ANOVA), and the result is shown in Figure 5.5.

ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>D f</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	3925.96	1	3925.96	6.70720	0.02695	4.96460
Within Groups	5853.35	1	585.335	4873	7951	2744
	3669	0	3669			
	9779.31	1				
Total	7895	1				

Figure 5.5: Results of ANOVA.

From Figure 5.5, the result shows that the proposed method scores are significantly different from the existing method under the non-stationary environment. This further reveals that the proposed method has a better and statistically significant result. The p-value selected is 0.05.

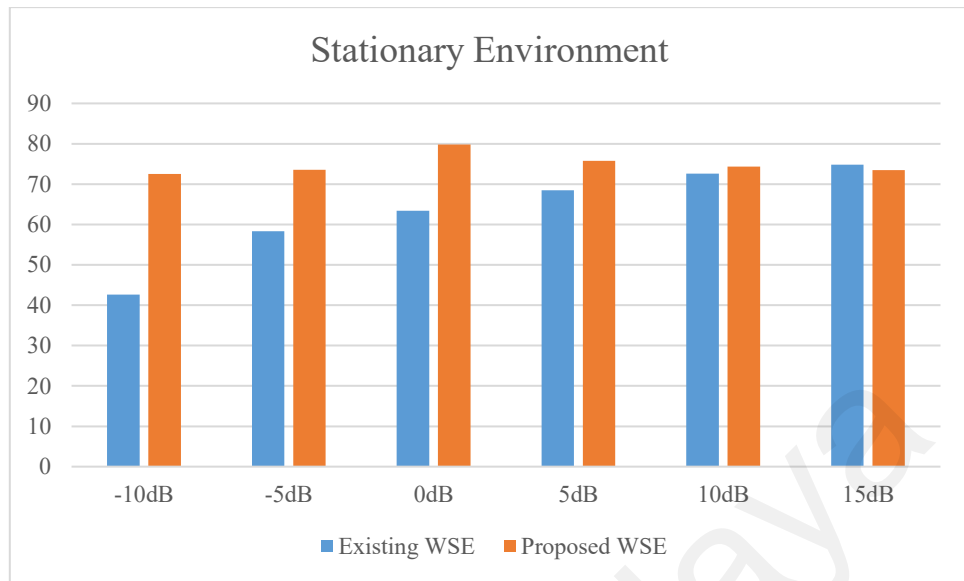


Figure 5.6: WRR of Existing Vs Proposed MCSE under non stationary environment

Figure 5.6 depicts the differences in the WRR for both the existing MCSE and proposed MCSE under stationary noises. The proposed MCSE is very effective for recognizing speech in stationary noisy environments compare to existing MCSE. Finally, to determine whether the result for existing MCSE and proposed MCSE was significantly different, Analysis of Variance (ANOVA) was conducted, and the result is shown in Figure 5.5.

ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>D f</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	400.092	1	400.092	5.45322	0.04167	4.96460
Within Groups	733.679	9	81.5199	3	8	3
Total	1133.77	10				

Figure 5.7: Results of ANOVA.

From Figure 5.7, the result indicates that the proposed method scores are significantly different from the existing method under stationary environment. This shows that proposed method has a better and statistically significant result. The p-value selected is 0.05.

5.5 Summary

This research proposed the noise filtering framework in MCSE using DWT and CNN-BLSTM algorithms. The proposed MCSE system gives better results than the existing MCSE system under noisy environments from -10db to 15db levels of SNR. It proved that the proposed feature selection and deep learning algorithms performed well at low SNR's for MCSE under noisy environments.

CHAPTER 6: CONCLUSION AND FUTURE WORKS

6.1 Overview

This research aimed to propose a noise filtering framework using suitable algorithm(s) for Multi-Channel Speech Enhancement systems in handling various Signal-to-Noise ratio (SNR) of environmental noises. This chapter summarizes the work that was carried out in this research. The research objectives listed in chapter one are revisited, also, the research contributions, some limitations of this research, and some suggestions for future works are discussed in the following sections.

6.2 Fulfilment of Research Objectives

This section discusses the accomplishments of the research objectives defined for this research.

6.2.1 Research Objective 1

The first objective is to analyse the existing speech enhancement systems in handling different types of noises. This objective is achieved with the analysis of the findings of existing algorithms and components involved in the speech enhancement system in section 2.3 of chapter 2, reported in Table 2.1. and, it was achieved with the analysis of the findings of existing algorithms and components involved in the Multi-Channel Speech Enhancement system in section 2.4 of chapter 2, reported in Table 2.2.

The findings describe the speech enhancement, multi-channel speech enhancement algorithms, deep learning approaches in speech enhancement and the performance of these algorithms in terms of speech quality under different noises. algorithms. The findings also

described preprocessing algorithms and speech classification algorithms in sections 2.5.1 and 2.5.2.

The first research question was answered as follow:

RQ1: What are the components, algorithms used and the existing of the speech enhancement systems in handling environmental noises?

Section 2 and Table 2.1 summarized the algorithms involved in speech enhancement based on deep learning methods and their performance in improving the speech quality. The conventional speech enhancement such as spectral subtraction, Wiener filtering, and minimum mean square error, have been outperformed by deep learning methods. CNN has the capacity to detect patterns in neighbouring speech structures. Compared to RNN (8.31% WER on Aurora-4) and standard DNN (14.7% WER on CHiME-2, CNN (6.3% WER on CHiME-4 and AURORA database) is more effective.

Section 2.4, section 2.4.3 and Table 2.2 summarized the components and algorithms in Multi-Channel Speech Enhancement system. Beamforming, adaptive noise reduction, voice activity detection and their performance under different types of noises were analysed.

From this objective, the research analysed the existing MCSE and identified the limitations of Multi-Channel Speech Enhancement system.

6.2.2 Research Objective 2

The second objective is to experiment the performance of the existing Multi-Channel Speech Enhancement systems in handling environmental noises. This research conducted benchmark experiments on MCSE to identify the real problem existing in MCSE.

This objective presented the experimental study of existing Multi-Channel Speech Enhancement system where the existing schemes such as beamforming, adaptive noise cancellation, and voice activity detection were described in section 3.4. It also presented the experimental design and experimental setup in 3.5.1 and 3.5.2 and reported in Table 3.2 accordingly. Moreover, it presented the outcome of the implementation of VAD (Voice Activity Detection), ANR (Adaptive Noise Reduction) and beamforming algorithms for MCSE in sections 3.7.1 and 3.7.2. Finally, it compared the MCSE systems under stationary and non-stationary noisy environments in Figure 3.17 and 3.18.

The second research question was answered as follow:

RQ2: What is the performance of existing Multi-Channel Speech Enhancement systems under noisy environments?

From the benchmark experiments, this research has identified that the MCSE's recognition rate reported the highest WRR at 93.77% for high SNR (at 20dB) and 5.64% for low SNR (at -10dB) on an average of five types of different noises (Pavani Cherukuru et al., 2021). Based on the benchmark experiment results, existing MCSE is giving an acceptable recognition rate at high SNR but the results are not acceptable at low SNR's, so there is a need to solve this problem to improve recognition rate at low SNR's.

6.2.3 Research Objective 3

The third research objective is to develop a noise-filtering framework using suitable algorithm(s) in Multi-Channel Speech Enhancement systems for filtering various Signal-to-Noise ratio (SNR) of environmental noises. This objective is achieved by using a sequence of experiments as discussed in sections 4.2.1 and 4.2.2 of chapter 4. In Figure 4.1, the overall system development is depicted. Discrete wavelet transforms and CNN-BLSTM algorithms are explained in detail in sections 4.2.1 and 4.2.2 and reported in Table 4.1 and Table 4.2 in chapter 4. The dataset details, experimental design, experimental setup and software requirements are explained in sections 4.3, 4.4 and 4.5 respectively.

The third research question was answered as follow:

RQ3: Which algorithms can be suitable to apply on the proposed MCSE framework in improving the performance at various Signal-to-Noise ratio (SNR) of environmental noises? The existing research focused on improving the performance of speech communication devices. Several algorithms were presented to improve the speech quality in MEMS microphones, but these algorithms suffer from low performance. To overcome this problem, this research proposed noise filtering framework using preprocessing and deep learning algorithms.

Based on (Ping et al., 2019, Katti et al., 2011 and Maria Labied et al., 2021), DWT preprocessing algorithm and CNN algorithm are the suitable algorithms to filter noisy environments and improve the quality of speech as reported in Table 1 and section 2.5.1. According to Maria Labied et al., 2021 Discrete wavelet transform is effective in denoising

speech signals, and it can compress the speech signal without degrading the speech quality. However, there is a limitation in computational complexity which is not flexible. Kattia et al., 2011 and Ping et al., 2019 stated that CNN has the capacity to detect patterns in neighbouring speech structures and compared to RNN and standard DNN, CNN is more effective. However, it has inability to maintain invariance when the input data changes. Among the deep learning algorithms, CNN reported the highest word recognition rate (WRR) at 90.45%, and the lowest WRR at 87.45% on an average for environmental noises (Pavani cherukuru et al., 2021).

The proposed framework in MCSE involved with preprocessing algorithm based on DWT and deep learning algorithm based on CNN achieved better performance in terms of word recognition rate as compared to the existing algorithms such as Beamforming, Adaptive noise reduction and Voice activity detection under noisy environments especially at low SNR's which is better than existing MCSE system in reference to word recognition rate accuracy.

6.2.4 Research Objective 4

The fourth objective is to evaluate the performance of the developed noise filtering framework in Multi-Channel Speech Enhancement system in handling various Signal-to-Noise ratio (SNR) of environmental noises.

The evaluation measurements used to evaluate the proposed framework are spectrogram analysis and word recognition rate (WRR) explained in section 5.2.1 and 5.2.2. The performance measure parameters are also used to test the performance of noise filtering framework as explained in 5.2.3.

The results of proposed framework are analysed in section 5.3, detailed spectrogram analysis of noisy speech signal and enhanced speech signals under airport, babble, restaurant and white gaussian noises are described at different levels of SNR in section 5.3.1 and reported in Table 5.2, 5.3, 5.4 and 5.5 respectively. The results of word recognition rate of the proposed noise filtering framework is explained in section 5.3.2 and reported in Table 5.6 which is 70.55% of lowest WRR at -10db and 78.75% of highest WRR at 0db in airport noise, 68.55% of lowest WRR at -10db and 73.79% of highest WRR at 10db in babble noise, 72.50% of lowest WRR at -10db and 81.49% of highest WRR at 10db in car noise, 65.15% of lowest WRR at -10db and 80.45% of highest WRR at 0db in exhibition noise, 72.50% of lowest WRR at -10db and 75.75% of highest WRR at 0db in restaurant noise and with stationary noise which is gaussian noise, resulted at 72.51% of lowest WRR at -10db and 79.78 of highest WRR at 0db SNR. Also, the performance measurement parameters of proposed noise filtering framework is explained in section 5.3.3 and reported in in Table 5.7, 5.8, 5.9 and 5.10 respectively under airport, babble, restaurant and gaussian noises.

The fourth research question is answered as follow:

RQ4: Can the proposed research improve the performance of the MCSE in filtering environmental noises with acceptable results?

The performance of the developed noise filtering framework in handling various SNR of environmental noises show a WRR of 70.55% at -10dB SNR and 75.44 % at 15dB SNR, while 5.82 % at -10dB and 88.8% at 15dB by the existing MCSE system. It proved that the proposed feature selection and deep learning algorithms performed well at low SNR's for MCSE under noisy environments.

The proposed system is therefore considered as an effective solution as it obtained considerable performance among all the related studies. Section 5.4 showed a comparison of this study with related existing studies with regards to the proposed system. By comparing the performance of the developed noise filtering framework in handling various SNR of environmental noises, it achieved a WRR of 70.55% at -10dB SNR and 75.44 % at 15dB SNR, while 5.82 % at -10dB and 88.8% at 15dB by the existing MCSE system. It can be inferred from the comparison that the proposed system performed well when compared with other related studies.

From the ANOVA analysis, the result indicated that the proposed method scores are significantly different from the existing method. This proves that proposed method has a better and statistically significant result.

6.3 Research Contributions

The current research contributes to the field of Multi-Channel Speech Enhancement system by improving its recognition accuracy by applying the combined preprocessing algorithm based on discrete wavelet transform and CNN- BLSTM algorithms. The contribution of this study can be listed as follows:

- Development of noise filtering framework for Multi-Channel Speech Enhancement system to improve the recognition accuracy.
- The proposed method helps to filter the noises in speech from levels of SNR to high levels of SNR which mostly enhance the Multi-Channel Speech Enhancement performance rate.

- Investigate preprocessing algorithms and its performance in speech quality from the literature review. The performance in speech quality helps to identify the suitable preprocessing algorithm for noise filtering framework.
- The proposed framework could help the MCSE system to be used in any outdoor environments.
- The new knowledge that been added to this research area is multi-channel speech enhance system and is capable to filter low SNR noises under noisy conditions.

6.4 Research Limitation:

- This research focuses on multi-channel speech enhancement systems which is used in speech communication devices that incorporate the speech recognition system. It can filter environmental noises but does not include other noises.
- In both the preliminary and proposed experiments, the testing and training stage, the AURORA speech noisy databases were used. Results may vary with different dataset.
- Computation level is high.

6.5 Suggestions for Future Works

This section provides some suggestions from carrying out this research, which can help to enhance the performance of the Multi-Channel Speech Enhancement system.

- Improve this proposed framework to be included in assistive technology to help motor skill impaired people to improve their socio communication skills for a better quality of life.
- The current research on speech processing has been focused on deep learning algorithms. Notably, the deep learning-based speech enhancement method can outperform the existing methods. Nevertheless, this approach involves substantially higher computing costs. Thus, it is difficult to implement deep learning-based approaches in portable communication devices that require a low computing complexity for real-world implementations. In this regard, it may be desirable to combine the existing filters and deep learning approaches to enhance the performance in terms of both the speech quality and intelligibility.
- Conducting more investigation of all kinds of noises, speech enhancement algorithms and its effectiveness which can help to make the complete noise free Multi-Channel Speech Enhancement system.
- Apply more combination of pre-processing and complicated classification algorithms like deep learning algorithms to enhance the recognition accuracy.
- Enhance the proposed framework to be applied in wearable technology applications to be used in real world environments. Thus, the motor skill is impaired to help improve the socio communication in the real world.

References:

- A. Katti and S. K. Anusuya, "Front end analysis of speech recognition : a review," *Int. J. Speech Technol.*, vol. 14, no. 2, pp. 99– 145, 2011, doi: 10.1007/s10772-010-9088-7.
- Abdel-Hamid, A. Mohamed, H. Jiang and G. Penn, Applying convolutional neural networks concepts to hybrid NNHMM model for speech recognition, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4277–4280, 2012.
- Abel, J., Kaniewska, M., Guillaum , C., Tirry, W., &Fingscheidt, T. (2016). An instrumental quality measure for artificially bandwidth-extended speech signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(2), 384-396.
- Adeel, A., Ahmad, J., Larijani, H., & Hussain, A. (2020). A novel real-time, lightweight chaotic-encryption scheme for next-generation audio-visual hearing aids. *Cognitive Computation*, 12(3), 589-601.
- Alexandre, E.,  lvarez, L., Cuadra, L., Rosa-Zurera, M., &Vicen-Bueno, R. (2008, May). A constructive algorithm for multilayer perceptrons for speech/non-speech classification in hearing aids. In *Audio Engineering Society Convention 124*. Audio Engineering Society.
- Ali, Hazrat, et al. "DWT features performance analysis for automatic speech recognition of Urdu." *SpringerPlus* 3.1 (2014): 1-10.
- Ananda Krishna B. & Yadav G.V.P.C. (2016). Performance comparison of different variable filters for noise cancellation in real-time environment. *International Journal of Signal Processing, Image Processing and Pattern Recognition* 9(2): 107–126.
- Andersen, K. T., &Moonen, M. (2017). Robust speech-distortion weighted interframe Wiener filters for single-channel noise reduction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1), 97-107.

- Araki, S., Sawada, H., & Makino, S. (2007, April). Blind speech separation in a meeting situation with maximum SNR beamformers. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 (Vol. 1, pp. I-41). IEEE.
- Avalos, J. G., Sanchez, J. C., & Velazquez, J. (2011). Applications of adaptive filtering. *Adaptive Filtering Applications*, 1, 3-20.
- B.D. Van Veen, and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, 1988, pp. 4-24.
- Babu, G. Ramesh, et al. "Speech enhancement using beamforming." *Int. J. Eng. Comput. Sci* 4.4 (2015): 11143-11147.
- Babu, G. Ramesh, et al. "Speech enhancement using beamforming." *Int. J. Eng. Comput. Sci* 4.4 (2015): 11143-11147.
- Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in 2019 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc., 2019.
- Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal", *American Society for Engineering Education (ASEE) Zone Conference Proceedings*, 2008, pp. 1-7.
- Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal", *American Society for Engineering Education (ASEE) Zone Conference Proceedings*, 2008, pp. 1-7.
- Bactor, P., & Garg, A. (2012). Different Algorithms for the Enhancement of the Intelligibility of a Speech Signal. *International Journal of Engineering Research and Development*, 2(2), 57-64.

- Bactor, P., & Garg, A. (2012). Different Algorithms for the Enhancement of the Intelligibility of a Speech Signal. *International Journal of Engineering Research and Development*, 2(2), 57-64.
- Baghel, S., Prasanna, S. M., & Guha, P. (2020). Exploration of excitation source information for shouted and normal speech classification. *The Journal of the Acoustical Society of America*, 147(2), 1250-1261.
- Basu, Sumit, Steve Schwartz, and Alex Pentland. "Wearable phased arrays for sound localization and enhancement." *Digest of Papers. Fourth International Symposium on Wearable Computers*. IEEE, 2000.
- Benesty, J., & Jingdong, C. (2012). *Study and design of differential microphone arrays (Vol. 6)*. Springer Science & Business Media.
- Bertrand, A. (2011, November). Applications and trends in wireless acoustic sensor networks: A signal processing perspective. In *2011 18th IEEE symposium on communications and vehicular technology in the Benelux (SCVT)* (pp. 1-6). IEEE.
- Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.
- Buechner, A., Dyballa, K. H., Hehrmann, P., Fredelake, S., & Lenarz, T. (2014). Advanced beamformers for cochlear implant users: acute measurement of speech perception in challenging listening conditions. *PloS one*, 9(4), e95542.
- C Donahue, B.Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in 2018 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2018.
- C Donahue, B.Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in 2018 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2018.

- Chen, “Fundamentals of Noise Reduction in Spring Handbook of Speech Processing”, Springer, 2008.
- Chen, H., Abhayapala, T. D., & Zhang, W. (2015). Theory and design of compact hybrid microphone arrays on two-dimensional planes for three-dimensional soundfield analysis. *The Journal of the Acoustical Society of America*, 138(5), 3081-3092.
- Cherukuru, Pavani, Mumtaz Begum Mustafa, and Hema Subramaniam. "The Performance of Multi-Channel Speech Enhancement System Under Noisy Environment: An Experimental Study." *IEEE Access* 10 (2021): 5647-5659.
- Chiluveru, S. R., & Tripathy, M. (2021). Speech enhancement using a variable level decomposition dwt. *National Academy Science Letters*, 44, 239-242.
- Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., ... & R Cowan, B. (2019). The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers*, 31(4), 349-371.
- D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," in 2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc., 2018.
- Darabkh, K. A., Haddad, L., Sweidan, S. Z., Hawa, M., Saifan, R., & Alnabelsi, S. H. (2018). An efficient speech recognition system for arm-disabled students based on isolated words. *Computer Applications in Engineering Education*, 26(2), 285-301.
- Das, N., Chakraborty, S., Chaki, J., Padhy, N., & Dey, N. (2021). Fundamentals, present and future perspectives of speech enhancement. *International Journal of Speech Technology*, 24(4), 883-901.
- Dash, T. K., & Solanki, S. S. (2019). Insight on the utilization of the noise estimation algorithms in phase aware modified multiband spectral subtraction. *Australian Journal of Electrical and Electronics Engineering*, 16(4), 250-255.

- de Cheveigne, A. (2010). Time-shift denoising source separation. *Journal of Neuroscience Methods*, 189(1), 113-120.
- De Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4), 1917-1930.
- Dey, N., & Ashour, A. S. (2018). *Direction of arrival estimation and localization of multi-speech sources*. Berlin: Springer International Publishing.
- Elko, G. W. (2004). Differential microphone arrays. In *Audio signal processing for next-generation multimedia communication systems* (pp. 11-65). Springer, Boston, MA.
- Elko, G. W., & Meyer, J. (2008). Microphone arrays. In *Springer handbook of speech processing* (pp. 1021-1041). Springer, Berlin, Heidelberg.
- Ephraim, Yariv, Hanoch Lev-Ari, and William JJ Roberts. "A brief survey of speech enhancement 1." *Microelectronics* (2018): 20-1.
- Ezzine, H. Satori, M. Hamidi, and K. Satori, "Moroccan Dialect Speech Recognition System Based on CMU SphinxTools," in 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), Jun. 2020, pp. 1–5, doi: 10.1109/ISCV49265.2020.9204250.
- F. Li and S.-C. Chang, "Speech recognition of mandarin syllables using both linear predict coding cepstra and Mel frequency cepstra," in *Proceedings of the 19th Conference on Computational Linguistics and Speech Processing*, 2007, pp. 379–390, [Online]. Available: <https://www.aclweb.org/anthology/O07-2009/>.
- Feng, Y., & Chen, F. (2022). Nonintrusive objective measurement of speech intelligibility: A review of methodology. *Biomedical Signal Processing and Control*, 71, 103204.

- Fischer, D., & Doclo, S. (2018, September). Robust constrained MFMVDR filtering for single-microphone speech enhancement. In 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC) (pp. 41-45). IEEE.
- Flanagan, F. J. "1972 values for international geochemical reference samples." *Geochimica et Cosmochimica Acta* 37.5 (1973): 1189-1200.
- Furui, S. (2018). *Digital Speech Processing, Synthesis, and Recognition: Synthesis, and Recognition*.
- G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *Proc. Annu. Conf. Speech Communication Association Interspeech 2019*, 2019.
- Gabbay, A., Ephrat, A., Halperin, T., & Peleg, S. (2018, April). Seeing through noise: Visually driven speaker separation and enhancement. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3051-3055). IEEE.
- Galindo, M. B. (2020). *Microphone array beamforming for spatial audio object capture* (Doctoral dissertation, University of Surrey).
- Gannot, S., & Cohen, I. (2008). Adaptive beamforming and postfiltering. In *Springer handbook of speech processing* (pp. 945-978). Springer, Berlin, Heidelberg
- Grama and C. Rusu, "Audio signal classification using Linear Predictive Coding and Random Forests," in 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Jul. 2017, pp. 1-9, doi: 10.1109/SPED.2017.7990431.
- H. Veisi and H. Sameti, "The integration of principal component analysis and cepstral mean subtraction in parallel model combination for robust speech recognition," *Digit. Signal Process.*, vol. 21, no. 1, pp. 36-53, Jan. 2011, doi: 10.1016/j.dsp.2010.07.004.

Hansler, E., & Schmidt, G. (Eds.). (2008). *Speech and audio processing in adverse environments*. Springer Science & Business Media.

Haykin, Simon, et al. "Classification of radar clutter in an air traffic control environment." *Proceedings of the IEEE* 79.6 (1991): 742-772.

[https://ecs.utdallas.edu/loizou/speech/noizeus/#:~:text=A%20noisy%20speech%20corpus%20\(NOIZEUS,world%20noises%20at%20different%20SNRs.](https://ecs.utdallas.edu/loizou/speech/noizeus/#:~:text=A%20noisy%20speech%20corpus%20(NOIZEUS,world%20noises%20at%20different%20SNRs.)

Hu, Yi, and Philipos C. Loizou. "Subjective comparison of speech enhancement algorithms." 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. Vol. 1. IEEE, 2006.

Huang, G., Chen, J., & Benesty, J. (2019). Design of planar differential microphone arrays with fractional orders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 116-130.

J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville and Y. Bengio, Maxout networks, in: **Proceedings of the 30th International Conference on Machine Learning, Proceedings of Machine Learning Research**, pp. 1319–1327, 2013.

J. Rownicka, P. Bell, and S. Renals, "Multi-Scale octave convolutions for robust speech recognition," in *IEEE Int. Conf. Acoustics Speech and Signal Processing Proc.*, 2020.

J.-M. Valin, "On adjusting the learning rate in frequency domain echo cancellation with double-talk," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1030–1034, 2007.

J.-S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 2, pp. 373–376, 1990.

J.-Y. Lee and J. Hung, "Exploiting principal component analysis in modulation spectrum enhancement for robust speech recognition," in *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Jul. 2011, pp. 1947–1951, doi: 10.1109/FSKD.2011.6019893.

Johnson, Don H. "Signal-to-noise ratio." *Scholarpedia* 1.12 (2006): 2088.

Johnson, L., Adams Becker, S., Cummins, M., and Estrada, V. (2013). *Technology Outlook for Norwegian Schools 2013-2018: An NMC Horizon Project Regional Analysis*. Austin, Texas: The New Media Consortium.

Jokinen, Emma, et al. "Enhancement of speech intelligibility in near-end noise conditions with phase modification." *Fifteenth Annual Conference of the International Speech Communication Association*. 2014.

K. He, X. Zhang, S. Ren and J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Kang, T. G., Shin, J. W., & Kim, N. S. (2018). DNN-based monaural speech enhancement with temporal and spectral variations equalization. *Digital Signal Processing*, 74, 102-110.

Kanisha, B., Lokesh, S., Kumar, P. M., Parthasarathy, P., & Babu, G. C. (2018). Speech recognition with improved support vector machine using dual classifiers and cross fitness validation. *Personal and ubiquitous computing*, 22(5), 1083-1091.

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., ... & Zhang, W. (2019, December). A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 449-456). IEEE.

Karjol, M. A. Kumar, and P. K. Ghosh, "Speech enhancement using multiple deep neural networks," in *2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc.*, 2018.

Karthik, A., & MazherIqbal, J. L. (2021). Efficient Speech Enhancement Using Recurrent Convolution Encoder and Decoder. *Wireless Personal Communications*, 1-15.

- Ken, Chen, Huang Wei, and Wang Min. "Wearable support system for intelligent workshop application." Computational Problem-Solving (ICCP), 2012 International Conference on. IEEE, 2012.
- Kinoshita, T.Ochiai, M.Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in 2020 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc. 2020.
- Kinoshita, T.Ochiai, M.Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in 2020 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc. 2020.
- Kokkinakis, Kostas, and Philipos C. Loizou. "Advances in modern blind signal separation algorithms: theory and applications." Synthesis lectures on algorithms and software in engineering 2.1 (2010): 1-100.
- Krishna, B. Ananda, and GVP Chandra Sekhar Yadav. "Performance comparison of different variable filters for noise cancellation in real-time environment." International Journal of Signal Processing, Image Processing and Pattern Recognition 9.2 (2016): 107-126.
- Kumatani, K., Minhua, W., Sundaram, S., Ström, N., & Hoffmeister, B. (2019, May). Multi-geometry spatial acoustic modeling for distant speech recognition. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6635-6639). IEEE.
- L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in 13th Annu. Conf. Speech Communication Association Interspeech 2012, 2012.
- L. Toth, Convolutional deep maxout networks for phone recognition, **Interspeech**, pp. 1078–1082, Singapore, September 14–18, 2014.

- Labied, Maria, and AbdessamadBelangour. "Automatic Speech Recognition Features Extraction Algorithms: A Multi-criteria Comparison." *International Journal of Advanced Computer Science and Applications* 12.8 (2021).
- Labied, Maria, and AbdessamadBelangour. "Automatic Speech Recognition Features Extraction Algorithms: A Multi-criteria Comparison." *International Journal of Advanced Computer Science and Applications* 12.8 (2021).
- Lee, M., Lee, J., & Chang, J. H. (2019). Ensemble of jointly trained deep neural network-based acoustic models for reverberant speech recognition. *Digital Signal Processing*, 85, 1-9.
- Leman A, Faure J and Parizet E, "Influence of informational content of background noise on speech quality evaluation for VoIP application", *The Journal of the Acoustical Society of America* 123(5):3066, 2008.
- Levin, D., Habets, E. A., &Gannot, S. (2012). Maximum likelihood estimation of direction of arrival using an acoustic vector-sensor. *The Journal of the Acoustical Society of America*, 131(2), 1240-1248.
- Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014). Auditory feedback of one's own voice is used for high-level semantic monitoring: the "self-comprehension" hypothesis. *Frontiers in human neuroscience*, 8, 166.
- Liu, X., & Pons, J. (2021, June). On permutation invariant training for speech source separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6-10). IEEE.
- M. H. Soni, N. Shah, and H. A. Patil, "Time-Frequency masking-based speech enhancement using generative adversarial network," in *2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc.*, 2018.

- M. Iosif, G. Todor, S. Mihalis, and F. Nikos, "Comparison of Speech Features on the Speech Recognition Task," J. Comput. Sci., vol. 3, no. 8, pp. 608–616, 2007.
- M. Wollmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," in 2013 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2013, pp. 6822-6826.
- Malik, O. P., Hope, G. S., & Cheng, S. J. (1991). Some issues on the practical use of recursive least squares identification in self-tuning control. International Journal of Control, 53(5), 1021-1033.
- Malik, O. P., Hope, G. S., & Cheng, S. J. (1991). Some issues on the practical use of recursive least squares identification in self-tuning control. International Journal of Control, 53(5), 1021-1033.
- Megan Rose Dickey, (2013). www.go.nmc.org/nex, <http://www.businessinsider.com/wearable-tech-is-the-next-big-thing>.
- Meyer, J., & Elko, G. (2008, March). Spherical harmonic modal beamforming for an augmented circular microphone array. In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 5280-5283). IEEE.
- Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. European Journal of Neuroscience, 48(7), 2609-2621.
- Miyazaki, R. (2019). Establishing document creation support for people with upper limb disabilities by hands-free speech recognition and gaze detection. Impact, 2019(10), 67-69.

- Mohammad A Akhaee, Ali Ameri and Farokh A Marvasti, "Speech Enhancement by Adaptive Noise Cancellation in the Wavelet Domain" 5 th International Conference on Information Communications & Signal Processing, 2005.
- Mokbel, Chafic, Denis Jouvét, and Jean Monné. "Deconvolution of telephone line effects for speech recognition." *Speech communication* 19.3 (1996): 185-196.
- Moquin, P. (2004). *Beamforming using scattering conformal microphone arrays* (Doctoral dissertation, Carleton University).
- Mouaz, B. H. Abderrahim, and E. Abdelmajid, "Speech recognition of Moroccan dialect using hidden Markov models," in *Procedia Computer Science*, Jan. 2019, vol. 151, pp. 985–991, doi: 10.1016/j.procs.2019.04.138.
- Naik, G. R. (2012). Measure of quality of source separation for sub-and super-Gaussian audio mixtures. *Informatica*, 23(4), 581-599.
- Narine, M. (2020). Active noise cancellation of drone propeller noise through waveform approximation and pitch-shifting.
- Nongpiur, R. C., and D. J. Shpak. "Impulse-noise suppression in speech using the stationary wavelet transform." *The Journal of the Acoustical Society of America* 133.2 (2013): 866-879.
- Novotný, O., Plchot, O., Glembek, O., & Burget, L. (2019). Analysis of DNN speech signal enhancement for robust speaker recognition. *Computer Speech & Language*, 58, 403-421.
- Oliinyk, V., Lukin, V., & Djurovic, I. (2020, October). A Fast and Efficient Method for Time Delay Estimation for the Wideband Signals in Non-gaussian Environment. In *Conference on Integrated Computer Technologies in Mechanical Engineering– Synergetic Engineering* (pp. 30-41). Springer, Cham.

- P. Karjol, M. A. Kumar, and P. K. Ghosh, "Speech enhancement using multiple deep neural networks," in 2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc., 2018.
- P. Plantinga, D. Bagchi, and E. Fosler-Lussier, "An exploration of mimic architectures for residual network based spectral mapping," in 2019, 2019.
- P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in Proc. ICASSP, pp. 629- 632, 1996. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning – Data Mining, Inference, and Prediction. New York: Springer, 2009.
- P. Wang and D. L. Wang, "Enhanced spectral features for distortion-independent acoustic modeling," in Proc. Annu. Conf. Speech Communication Association Interspeech2019, 2019, pp. 476–480
- Page, Tom. "A forecast of the adoption of wearable technology", International Journal of Technology Diffusion (IJTD) 6.2 (2015): 12-29.
- Palla, Alessandro, et al. "Multi-Channel Speech Enhancement System for Motor Impaired People." International Conference on Applications in Electronics Pervading Industry, Environment and Society. Springer, Cham, 2017.
- Palla, Alessandro, et al. "Multi-Channel Speech Enhancement system based on MEMS microphone array for disabled people." Design & Technology of Integrated Systems in Nanoscale Era (DTIS), 2015 10th International Conference on. IEEE, 2015.
- Palla, Alessandro, et al. "Multi-Channel Speech Enhancement System for Motor Impaired People." International Conference on Applications in Electronics Pervading Industry, Environment and Society. Springer, Cham, 2017.

- Palla, Alessandro, et al. "Multi-Channel Speech Enhancement system based on MEMS microphone array for disabled people." Design & Technology of Integrated Systems in Nanoscale Era (DTIS), 2015 10th International Conference on. IEEE, 2015.
- Pandey and D. Wang, "A new framework for CNN-Based speech enhancement in the time domain," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 27, July. 2019.
- Park, Gyuseok, et al. "Speech enhancement for hearing aids with deep learning on environmental noises." Applied Sciences 10.17 (2020): 6077.
- Park, Se Rim, and Jinwon Lee. "A fully convolutional neural network for speech enhancement." arXiv preprint arXiv:1609.07132 (2016).
- Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in Proc. Annu. Conf. Int. Speech Communication Association Interspeech 2017, 2017.
- Pauline, S. H., Samiappan, D., & Kumar, R. B. (2021, July). Double Talk Detection in hands-free mobile communication-A comprehensive survey. In Journal of Physics: Conference Series (Vol. 1964, No. 6, p. 062044). IOP Publishing.
- Perry, T. L., Ohde, R. N., & Ashmead, D. H. (2001). The acoustic bases for gender identification from children's voices. The Journal of the Acoustical Society of America, 109(6), 2988-2998.
- Pinki Sahil Gupta" Speech Enhancement using spectral subtraction type algorithms: A Survey on Comparison" The IJCS Volume 04 Issue10-oct-2015,page no.1487-14878.
- Pradhan, G., & Prasanna, S. M. (2011, January). Significance of speaker information in wideband speech. In 2011 National Conference on Communications (NCC) (pp. 1-5). IEEE.
- Prasad, Ganga. "A review of different approaches of spectral subtraction algorithms for speech enhancement." Curr. Res. Eng 1.2 (2013): 57-64.

- R. L. K. Venkateswarlu and R. V. Kumari, "Novel approach for speech recognition by using self-Organized maps," in 2011 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), Apr. 2011, pp. 215–222, doi: 10.1109/ETNCC.2011.5958519.
- R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in Proc. Annu. Conf. Speech Communication Association Interspeech 2017, 2017.
- R. Venkatesha Prasad, A. Sangwan, H. Jamadagni, M. Chiranth, R. Sah, and V. Gaurav, "Comparison of voice activity detection algorithms for voip" , Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium on. IEEE, 2002, pp. 530–535.
- R. Venkatesha Prasad, A. Sangwan, H. Jamadagni, M. Chiranth, R. Sah, and V. Gaurav, "Comparison of voice activity detection algorithms for voip" , Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium on. IEEE, 2002, pp. 530–535.
- Rafaely, B. (2008). Spatial sampling and beamforming for spherical microphone arrays. 2008 Hands-Free Speech Communication and Microphone Arrays, 5-8.
- Rakesh, P., Priyanka, S. S., & Kumar, T. K. (2017, March). Performance evaluation of beamforming algorithms for speech enhancement. In 2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN) (pp. 1-5). IEEE.
- Ramana, A. V., LaxminarayanaParayitam, and Mythili Sharan Pala. "Investigation of automatic speech recognition performance and mean opinion scores for different standard speech and audio codecs." *IETE Journal of Research* 58.2 (2012): 121-129.
- Rao, Aparna, et al. "Neural correlates of selective attention with hearing aid use followed by ReadMyQuips auditory training program." *Ear and hearing* 38.1 (2017): 28-41.

- Rethage, J.Pons, and X. Serra, "A wavenet for speech p-ISSN: 1411-8289; e-ISSN: 2527-9955 denoising," in 2018 IEEE Int. Conf. IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2018.
- Reuven, G., Gannot, S., & Cohen, I. (2007). Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelobe canceller. *Speech communication*, 49(7-8), 623-635.
- Roß, B., Borgmann, C., Draganova, R., Roberts, L. E., & Pantev, C. (2000). A high-precision magnetoencephalographic study of human auditory steady-state responses to amplitude-modulated tones. *The Journal of the Acoustical Society of America*, 108(2), 679-691.
- Ryherd, E. E., Wayne, K. P., & Ljungkvist, L. (2008). Characterizing noise and perceived work environment in a neurological intensive care unit. *The Journal of the Acoustical Society of America*, 123(2), 747-756.
- S. Choi, J. H. Kim, J. Huh, A. Kim, J. W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Proc. Int. Conf. Learning Representations 2019*, 2019.
- S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Annu. Conf. Speech Communication Association Interspeech 2017*, 2017.
- S. W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. 9th Asia-Pacific Signal and Information Processing Association Annu. Summit and Conf. 2017*, 2018.
- Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2015-Janua, pp. 1468–1472, 2015.

- Santhanavijayan, A., Kumar, D. N., & Deepak, G. (2021). A semantic-aware strategy for automatic speech recognition incorporating deep learning models. In *Intelligent System Design* (pp. 247-254). Springer, Singapore.
- Saoud, S., Bennis, M., & Cherif, A. (2021). A modified speech denoising algorithm based on the continuous wavelet transformer and wiener filter. In *Innovative and Intelligent Technology-Based Services for Smart Environments–Smart Sensing and Artificial Intelligence* (pp. 119-126). CRC Press.
- Schroter, Hendrik, et al. "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.
- Seon Man Kin, "Wearable Hearing Device Spectral Enhancement Driven by Non-Negative Sparse Coding-Based Residual Noise Reduction", *Sensors Journals*, 2020.
- Seon Man Kin, "Wearable Hearing Device Spectral Enhancement Driven by Non-Negative Sparse Coding-Based Residual Noise Reduction", *Sensors Journals*, 2020.
- Shah, V. N. (2020). On variants of stochastic gradient descent (Doctoral dissertation).
- Shanmugapriya, N., and E. Chandra. "A thorough investigation on speech enhancement algorithms for hearing aids." *International Journal of Computer Applications* 99.13 (2014): 9-12.
- Siedenburg, K., Jacobsen, S., & Reuter, C. (2021). Spectral envelope position and shape in sustained musical instrument sounds. *The Journal of the Acoustical Society of America*, 149(6), 3715-3726.
- Skordilis, Z. I., et al. "Multichannel speech enhancement using MEMS microphones." *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

- Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, **J. Mach. Learn. Res.** 15 (2014), 1929–1958.
- Stanciu, Cristian, Lucian Stanciu, and Roxana Mihaescu. "Low Complexity Recursive Least-Squares Algorithm for Adaptive Noise Cancellation." ICN 2017 (2017): 113.
- Stupakov, Alex, et al. "The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments." *Computer Speech & Language* 26.1 (2012): 52-66.
- Sugiyama, A., Miyahara, R., & Oosugi, K. (2019, May). A noise robust hearable device with an adaptive noise canceller and its DSP implementation. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2722-2726). IEEE.
- Sun, Zhuoyi, et al. "A supervised speech enhancement method for smartphone-based binaural hearing aids." *IEEE Transactions on Biomedical Circuits and Systems* 14.5 (2020): 951-960.
- T. Gao, J. Du, L. R. Dai, and C. H. Lee, "Densely connected progressive learning for LSTM-Based speech enhancement," in 2018 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2018, pp. 5054-5058.
- T. Gao, J. Du, L. R. Dai, and C. H. Lee, "SNR-based progressive learning of deep neural network for speech enhancement," in Proc. Annu. Conf. Int. Speech Communication Association Interspeech, 2016.
- T. N. Sainath, A.-R. Mohamed, B. Kingsbury and B. Ramabhadran, Deep convolutional neural networks for LVCSR, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8614–8618, Vancouver, Canada, May 26–31, 2013.
- T. Takiguchi and Y. Ariki, "PCA-Based Speech Enhancement for Distorted Speech Recognition," *J. Multimed.*, vol. 2, no. 5, Sep. 2007, doi: 10.4304/jmm.2.5.13-18.

- Taal, Cees H., et al. "An algorithm for intelligibility prediction of time–frequency weighted noisy speech." *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011): 2125-2136.
- Taha, T. M., Adeel, A., & Hussain, A. (2018). A survey on algorithms for enhancing speech. *International Journal of Computer Applications*, 179(17), 1-14.
- Takaki, S., Kameoka, H., & Yamagishi, J. (2019). Training a neural speech waveform model using spectral losses of short-time fourier transform and continuous wavelet transform. *arXiv preprint arXiv:1903.12392*.
- Taseska, M., & Habets, E. A. (2014). Informed spatial filtering for sound extraction using distributed microphone arrays. *IEEE/ACM transactions on audio, speech, and language processing*, 22(7), 1195-1207.
- Tesch, K., Mohrmann, N. H., & Gerkmann, T. (2022). On the Role of Spatial, Spectral, and Temporal Processing for DNN-based Non-linear Multi-channel Speech Enhancement. *arXiv preprint arXiv:2206.11181*.
- Thomas, M. R. (2019, May). Practical concentric open sphere cardioid microphone array design for higher order sound field capture. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 666-670). IEEE
- Upadhyay, N., & Karmakar, A. (2015). Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study. *Procedia Computer Science*, 54, 574-584.
- van den Oord et al., "WaveNet: A generative model for raw audio," *arXiv*, 2016.
- Vaughan, R., & Andersen, J. B. (2003). Channels, propagation and antennas for mobile communications (No. 50). Iet.

- Vinay, H. C., Lavanya, P., Hippargi, A. A., Purohith, A., & Lohith, D. T. (2021, August). A Comparative Analysis on Speech Enhancement and Coding Algorithms. In 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT) (pp. 543-549). IEEE.
- Vincent, E., Virtanen, T., & Gannot, S. (Eds.). (2018). Audio source separation and speech enhancement. John Wiley & Sons.
- Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10), 1702-1726.
- Wang, Dong, Xiaodong Wang, and Shaohe Lv. "End-to-end mandarin speech recognition combining CNN and BLSTM." *Symmetry* 11.5 (2019): 644.
- Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich feature enhancement approach to the 2013 CHiME challenge using BLSTM recurrent neural networks," in 2nd Int. Workshop Machine Listening Multisource Environments, 2013.
- Widrow, J. R. Glover Jr, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong Jr, and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications", *Proceedings of the IEEE*, vol. 63, no. 12, pp. 1692– 1716, 1975.
- Willis, E. C., & Kenny, D. T. (2007). Variability in speaking fundamental frequency in the adolescent voice. *Proceedings of ICoMCS December 2007*, 172-175.
- Winursito, A., Hidayat, R., & Bejo, A. (2018, March). Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition. In 2018 International Conference on Information and Communications Technology (ICOIACT) (pp. 379-383). IEEE.
- Wu, D., Zhang, K., & Wei, Y. (2019, August). A Speech Enhancement System Based on Real-time Sound Source Localization and Super-directional Fixed Beamforming.

In 2019 IEEE International Conference on Real-time Computing and Robotics (RCAR) (pp. 334-339). IEEE.

- X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," 2014 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2014.
- X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in 2019 IEEE Workshop Applications Signal Processing to Audio and Acoustics, 2019.
- X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in Proc. Annu. Conf. Int. Speech Communication Association Interspeech 2013, 2013.
- Xu, Y., Yu, M., Zhang, S. X., Chen, L., Weng, C., Liu, J., & Yu, D. (2020). Neural spatio-temporal beamformer for target speech separation. arXiv preprint arXiv:2005.03889.
- Xu, Yong, et al. "Wearable microphone array as user interface." Proceedings of the fifth conference on Australasian user interface-Volume 28. Australian Computer Society, Inc., 2004.
- Xu, Yong, et al. "Wearable microphone array as user interface." Proceedings of the fifth conference on Australasian user interface-Volume 28. Australian Computer Society, Inc., 2004.
- Xu, Yong, et al. "Wearable microphone array as user interface." Proceedings of the fifth conference on Australasian user interface-Volume 28. Australian Computer Society, Inc., 2004.
- Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," in 2018 IEEE Int. Conf. Acoustics Speech and Signal Processing Proc., 2018, pp. 5074-5078.

- Yadava, T. G., & Jayanna, H. S. (2019). Speech enhancement by combining spectral subtraction and minimum mean square error-spectrum power estimator based on zero crossing. *International Journal of Speech Technology*, 22(3), 639-648.
- Yan, X., Yang, Z., Wang, T., & Guo, H. (2020). An Iterative Graph Spectral Subtraction Method for Speech Enhancement. *Speech Communication*, 123, 35-42
- Yoneyama, R., Wu, Y. C., & Toda, T. (2022). Unified Source-Filter GAN with Harmonic-plus-Noise Source Excitation Generation. arXiv preprint arXiv:2205.060
- Z. Ping, T. Li-Zhen, and X. Dong-Feng, "Speech Recognition Algorithm of Parallel Subband HMM Based on Wavelet Analysis and Neural Network," *Inf. Technol. J.*, vol. 8, no. 5, pp. 796–800, Jun. 2009, doi: 10.3923/itj.2009.796.800.
- Z. Xu, S. Elshamy, and T. Fingscheidt, "Using separate losses for speech and noise in mask-based speech enhancement," in 2020 IEEE Int. Conf. Acoustics, Speech and Signal Processing Proc., 2020.
- Zeiler, Matthew D., and Rob Fergus. "Stochastic pooling for regularization of deep convolutional neural networks." arXiv preprint arXiv:1301.3557 (2013).
- Zeng, Y., & Hendriks, R. C. (2013). Distributed delay and sum beamformer for speech enhancement via randomized gossip. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1), 260-273.
- Zhang, Y., Brooks, D. H., Franceschini, M. A., & Boas, D. A. (2005). Eigenvector-based spatial filtering for reduction of physiological interference in diffuse optical imaging. *Journal of biomedical optics*, 10(1), 011014.
- Zhao, Yan, et al. "DNN-based enhancement of noisy and reverberant speech." 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.