

ENHANCING SPEECH PERCEPTION IN COCHLEAR IMPLANTS: NOVEL APPROACHES IN ENCODING TEMPORAL FINE STRUCTURES AND NOISE REDUCTION

Thesis

Submitted in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

by

POLUBOINA VENKATESWARLU



DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
SURATHKAL, MANGALORE -575025

October, 2023

DECLARATION

I hereby *declare* that the research Thesis entitled **ENHANCING SPEECH PERCEPTION IN COCHLEAR IMPLANTS: NOVEL APPROACHES IN ENCODING TEMPORAL FINE STRUCTURES AND NOISE REDUCTION** which is being submitted to the *National Institute of Technology Karnataka, Surathkal* in partial fulfillment of the requirement for the award of the Degree of *Doctor of Philosophy* in Department of Electronics and Communication Engineering is a *bonafide report of the research work carried out by me*. The material contained in this research Thesis has not been submitted to any University or Institution for the award of any degree.



POLUBOINA VENKATESWARLU

Reg. No. 187067/187EC009

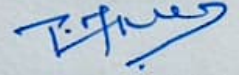
Department of Electronics and Communication Engineering.

Place: NITK-Surathkal.

Date: 11-10-2023

CERTIFICATE

This is to certify that the Research Thesis entitled **ENHANCING SPEECH PERCEPTION IN COCHLEAR IMPLANTS: NOVEL APPROACHES IN ENCODING TEMPORAL FINE STRUCTURES AND NOISE REDUCTION** submitted by **POLUBOINA VENKATESWARLU** (Register Number: 187EC009) as the record of the research work carried out by him, is accepted as the *Research Thesis submission* in partial fulfillment of the requirements for the award of degree of **Doctor of Philosophy**.



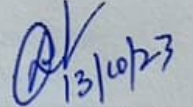
Dr. APARNA P

Research Guide

Assistant Professor

Dept. of Electronics and Communication Engg.

NITK Surathkal - 575025



Chairman-DRPC

प्राध्यापक एवं विभागाध्यक्ष / PROF & HEAD
(Signature with Date and Seal)

इ. एन. ए. सी. विभाग / Department of E & C

एन.आई.टी.के. सुल्तकल / NITK Surathkal

मंगलूर / MANGALURU - 575 025

Acknowledgements

Firstly, I would like to express my sincere gratitude and deep regards to my research adviser Dr. Aparna P. for her constant guidance and support provided throughout the course of this research work. Her patience, motivation and immense knowledge made me to accomplish this. I owe my deepest gratitude to her, also for bringing this thesis in the present form. I could not have imagined having a better adviser for my Ph.D.

Besides my adviser, I express my gratitude to Prof. Neelawar Shekar Vittal Shet, Head of the department, E&C Engg. for the constant support and encouragement. My sincere thanks goes to Prof. Laxminidhi T, Head, Dept. of E&C Engg. during my enrollment for Ph.D. program for his help and invaluable suggestions. Also, I thank Prof. Ashvini Chaturvedi, Head, Dept. of E&C Engg. during my research work for his support and help. My sincere thanks to Dr. Nagendrappa, Dr. Ragavendra Bobbi, RPAC members, for their invaluable suggestions during every step of my progress in the research work.

I want to thank Dr. Arivudai Nambi PitchaiMuthu for his invaluable assistance during my research and for providing me with a comprehensive explanation of cochlear implants. I am very grateful to Mr. Subramanya karanth for maintaining servers in MP&MC lab. I also take this opportunity to thank all the faculty and staff of E&C department, NITK Surathkal. I also thank the staff in Academic, Admission and Cash sections of NITK for their assistance.

I would like to express my gratitude to all friends and colleagues at NITK for encouraging me in good and bad times making a memorable stay in NITK. Thanks to Suresh N, Dr. Parthasaradi, Dr. Srinath Gunney, shajahan Abobeker, Sravan P, Balaram reddy, Vijay ratnam, Ashok chakravarthi, Sudhakar reddy, Dr. shareef babu for their support and help. Special thanks to my labmate, Dr. Lakshmi Poola, for her constant help during the initial stage of my research work. I also thank M.Tech. students- Srikanth M, Rohith kumar and Shesha sai who helped me in resolving technical issues.

I would like to thank Department of audiology and speech language pathology, Kasturba Medical College, Mangalore, Manipal Academy of Higher

Education, Manipal, India, for providing Auditory tools support during this project. Also, I express my gratitude to MHRD, for providing financial assistance under Teaching assistanship.

My heartfelt thanks to my lovely wife Mounika for her support and help throughout. In addition, my lovely sons, Mahadev Narayan, and second son are the source of bringing me immense joy throughout this journey. I recall with immense gratitude the unconditional love and support rendered by my family members and friends in my education and in my career throughout.

I am deeply indebted to all my teachers throughout the life, who have guided, encouraged and inspired me to grow in both technical and personal aspects. Finally, I would like to thank god for giving me good health, strength and bliss during my research work.

Dedicated to
My Dear Parents

Abstract

Cochlear implants (CIs) significantly enhance audibility and speech intelligibility in quiet environments. Nevertheless, speech recognition in noisy conditions remains a notable challenge. Efforts to enhance speech perception in cochlear implants typically follow two approaches: preprocessing, which involves improving the signal-to-noise ratio (SNR), and speech coding, aimed at encoding the significant cues necessary for speech recognition in noisy environments. The current thesis addresses both approaches. The initial approach involves encoding vital cues meaningfully, focusing on examining the impact of temporal fine structures through proportional frequency compression. In the second part, two denoising techniques are proposed as pre-processing to improve the SNR; one is the modified Wiener filter method, and the other one is the Deep denoising method for speech enhancement.

The research investigates the significance of TFS cut-off frequencies in CI speech coding to enhance speech perception in noise. Based on observations, an algorithm is introduced to represent TFS through proportionally frequency compressed cues. Additionally, a pitch-shifted overlap-add algorithm (PSOLA) is proposed to encode TFS within the neuro-physiological limitations of CI users. Speech recognition scores (SRS) are measured under various signal processing conditions, including a sinewave vocoder without TFS, four unshifted TFS conditions with varying frequency cut-offs, and three PSOLA conditions that shift TFS frequencies. The original envelope remains unchanged across all conditions. The results indicate that the SRS for TFS 600 Hz shifted to 300 Hz through PSOLA outperforms the no-TFS condition (sinewave vocoder), suggesting that encoding TFS using proportional frequency compression leads to improved speech perception in noise compared to the absence of TFS.

Furthermore, a modified Wiener filter method is proposed to enhance speech intelligibility specifically for noisy environments, focusing on the context of cochlear implants. This noise reduction technique aims to minimize the mean square error (MSE) between the temporal envelopes of the enhanced speech and the clean speech, making it suitable for CI appli-

cations. The study provides a theoretical analysis of the noise suppression function and evaluates its performance using objective and subjective tests. Objective measures such as the speech-to-reverberation modulation energy ratio (SRMR-CI) and extended short-time objective intelligibility (ESTOI) are employed, while subjective evaluation involves speech recognition through acoustic simulations of the cochlear implant. The proposed method’s performance is compared with the Wiener filter (WF) and sigmoidal functions, using the sinewave vocoder to simulate cochlear implant perception.

Finally, a new method is proposed for speech enhancement with deep learning training. The mathematical derivation supports the effectiveness of the proposed Noisy2Noisy_{avg} (N2N_{avg}) strategy over the Noise2Noise (N2N) strategy. The target and the input of a deep complex unit- network (DCU-Net) are trained solely using noisy speech samples, eliminating the need for a large number of clean speech samples. The proposed method is compared with state-of-the-art speech-denoising techniques. Experimental results demonstrate that the proposed approach not only reduces the reliance on clean targets but also mitigates the dependency on large data sizes typically associated with speech-denoising techniques.

In summary, this research addresses the limitations of current cochlear implant algorithms by proposing novel approaches for TFS encoding, noise reduction, and deep learning-based speech enhancement. The findings contribute to improving speech perception and intelligibility for individuals with cochlear implants, providing insights for further advancements in the field.

Keywords: Cochlear implants, TFS, Pitch shifting, Speech enhancement, Speech recognition, Deep speech denoising.

Contents

Acknowledgements	i
Abstract	v
List of figures	xi
List of tables	xiii
Abbreviations	xv
1 INTRODUCTION	1
1.1 Motivation	2
1.2 Objectives of the Thesis	4
1.3 Thesis Contributions	4
1.4 Thesis Organization	6
2 BACKGROUND	7
2.1 Human auditory system	7
2.1.1 Human auditory filters	9
2.2 Electric and Acoustic Stimulation	9
2.3 Commercially available cochlear implants	10
2.4 Signal Processing Strategies In Cochlear Implants	11
2.4.1 Continuous Interleaved Sampling	13
2.4.2 Spectral peak (SPEAK) Strategy	14
2.4.3 Advanced Combination Encoder (ACE)	14
2.4.4 Fine Structure Processing (FSP)	15
2.4.5 HiRes120	17
2.4.6 Single Side-band Encoder (SSE)	18
2.4.7 Harmonic Single Side-band Encoder (HSSE)	18
2.4.8 Temporal Limits Encoder (TLE)	19
2.5 Noise Reduction Methods for cochlear implants	19

2.5.1	Ideal Binary Mask	20
2.5.2	Wiener Filter	20
2.5.3	Sigmoidal function	20
2.5.4	Spectral Subtraction	21
2.5.5	Ideal Ratio Mask	21
3	DESIGN OF COCHLEAR ACOUSTIC MODEL TO ENCODE TEM- PORAL FINE STRUCTURES	23
3.1	Introduction	24
3.2	Investigating the significance of TFS for improving speech intelligibility	26
3.2.1	Experimental design and results	30
3.2.1.1	Quick Speech In Noise (QuickSIN)	30
3.2.1.2	Subjects	31
3.2.1.3	Stimuli	31
3.3	Encoding frequency compressed TFS within the neuro-physiological limitations (within 300Hz) of the CI users	33
3.3.1	TFS cut-off frequencies	33
3.3.2	Pitch synchronous overlap add algorithm (PSOLA) for TFS shifting	36
3.4	Experimental Results	40
3.4.1	Participants	40
3.4.2	Speech Recognition In Noise (SRIN) Test	41
3.4.3	Effect of TFS cut off frequency	44
3.4.4	Effect of TFS pitch shifting	45
3.5	Discussion	46
3.6	Summary	47
4	DESIGN OF AN EFFICIENT NOISE REDUCTION METHOD TO IMPROVE SPEECH RECOGNITION IN CIs	49
4.1	Introduction	50
4.1.1	Noise reduction methodology	51
4.2	The proposed noise suppression function	53
4.3	Processing Steps for proposed method	57
4.4	Simulation Results	59
4.4.1	Objective evaluation of the noise suppressed functions using MSE	62

4.4.2	Speech intelligibility of cochlear implants	63
4.4.3	Acoustic assessment of speech in noise	65
4.4.4	Subjective evaluations	68
4.5	Discussion	73
4.6	Summary	75
5	DEEP DENOISING METHOD FOR IMPROVING SPEECH RECOGNITION	77
5.1	Introduction	77
5.2	Proposed Method	79
5.2.1	DCUNET Architecture	82
5.2.2	Methodology for Training and Testing	84
5.3	Experimental results and discussion	85
5.3.1	Dataset Generation	85
5.3.2	Objective parameters	87
5.3.2.1	Signal to distortion ratio (SDR) measures	87
5.3.2.2	Perceptual evaluation of speech quality (PESQ)	87
5.3.2.3	Short time objective intelligibility (STOI)	87
5.3.2.4	Composite measures	88
5.3.3	Validation of $N2N_{avg}$ strategy	92
5.3.4	Evaluation of $N2N_{avg}$ with mismatch condition	95
5.4	Summary	96
6	CONCLUSIONS AND FUTURE WORK	97
6.1	Conclusions	97
6.2	Future Work	99
	Publications based on the thesis	113

List of Figures

1.1	Functional block diagram of a cochlear implant.	2
1.2	Block diagram of the cochlear implantation	2
2.1	Schematic diagram of the human auditory system	8
2.2	Responses of a sample of human auditory filters	9
2.3	Block diagram of the CIS stimulation strategy	13
2.4	Block diagram illustrating ACE.	14
2.5	Block diagram of FSP strategy	16
2.6	Generating pulses based on positive zero crossings	16
2.7	Block diagram of HiRes120 strategy	17
3.1	The block diagram of the three speech encoding methods in CIs	26
3.2	Filter bank response for 16 bands	27
3.3	SV Vs Full band TFS	32
3.4	Block diagram representation of TFS cut off frequencies	33
3.5	Block diagram representation of the proposed speech encoder with PSOLA	36
3.6	Pitch marking at local maxima of TFS	38
3.7	Psychometric plots of eight signal processing techniques	42
3.8	Speech intelligibility of sinewave vocoder and TFS cut-off frequencies	44
3.9	Speech intelligibility of sinewave vocoder and pitch shifted TFS	45
4.1	Block diagram representing noise reduction and vocoder-based simula- tion of cochlear implants	52
4.2	Comparison of the noise suppression functions with different a priori SNRs	56

4.3	Block diagram of steps involved in psychoacoustic studies and objective assessment.	58
4.4	Power spectrum of speech shape noise and 4-talker babble noise.	59
4.5	The waveform of (a) clean, (b) noisy, speech signals enhanced by (c) PM, (d) WF, and (e) Sigmoidal function. Spectrogram representation of (f) clean, (g) noisy, and speech signals enhanced by (h) PM, (i) WF, and (j) Sigmoidal function.	60
4.6	The expanded waveform of (a) clean, (b) noisy, speech signals enhanced by (c) PM, (d) WF, and (e) Sigmoidal function.	61
4.7	WF (blue) and PM (red) represent the mean square error concerning input SNR.	63
4.8	Speech intelligibility of CIs according to SRMR-CI metrics with different SNR levels.	64
4.9	ESTOI scores for speech signals corrupted by speech shape noise at different SNR levels.	66
4.10	ESTOI scores for speech signals corrupted by babble noise at different SNR levels.	67
4.11	The psychometric plots for the speech recognition ability of volunteers with speech shape noise.	70
4.12	The psychometric plots for the speech recognition ability of volunteers with babble noise.	72
5.1	Architecture of a ten-layer DCU-net for speech denoising.	82
5.2	Overview of training methodologies.	85
5.3	The training loss and validation loss.	88
5.4	Spectrograms of a speech database (Valentini-Botinhao et al. 2016) sentence: (a) clean speech, (b) noisy speech (babble noise at 0 dB), (PESQ = 1.623), (c) denoised speech by N2N (PESQ = 1.96) (d) enhanced speech by N2N _{avg} (PESQ = 2.00).	91
5.5	SDR comparison of noisy input signals and N2N _{avg} results.	93
5.6	MOS evaluation results from a subjective perspective.	93

List of Tables

2.1	A comparison of different coding strategies used in commercial CIs.	11
2.2	Summary of time frequency masking methods for noise reduction	22
3.1	Center frequencies of the corresponding channel number	28
3.2	Sample procedure of lists presented to NH Volunteers	41
3.3	Mean speech intelligibility scores (PC) of volunteers	43
3.4	Bayesian paired sample T-Test between Sinewave vocoder and TFS cut-off frequencies	44
3.5	Bayesian paired sample T-test between sinewave vocoder and pitch shifted TFS	46
4.1	Mean square error at different SNR levels with speech shape noise	62
4.2	Speech intelligibility of CIs according to SRMR-CI metrics with differ- ent SNR levels.	64
4.3	ESTOI values (D) for each noise reduction method with speech shape noise	66
4.4	ESTOI values (D) for each noise reduction method with babble noise	67
4.5	Normal hearing Participants details.	68
4.6	Mean speech intelligibility (PC) of volunteers with speech shape noise	69
4.7	The average SRTN of each noise-reduction method with speech shape noise	70
4.8	SRTN with speech shape noise	71
4.9	Quality assessment based on statistical tests of psychoacoustic experi- ments	71
4.10	Mean speech intelligibility (PC) of volunteers with babble noise	72
4.11	The average SRTN of each noise-reduction method with 4 talker babble noise	73

5.1	Literature review of training methods and Models used for speech enhancement.	78
5.2	Methodology for training and evaluation	85
5.3	Dataset generation	85
5.4	Performance comparisons of different methods in terms of STOI and PESQ with the noisy dataset I	89
5.5	Performance comparisons of N2C, N2N, and N2N _{avg} based on composite measures with the noisy dataset I	90
5.6	Result of VoiceBank-Demand (without mismatch in training and testing data).	92
5.7	Evaluation of training methods with SLR TELUGU dataset	95
5.8	Result of SLR66 (x_{lst2})+N ₁ (with a mismatch in training and testing data).	96

Abbreviations

Abbreviation	Expansion
ACE	Advanced Combination Encoder
CI	Cochlear Implant
CIS	Continuous Interleaved Sampling
DCU-Net	Deep Complex Unit Network
ENV	Envelope
FDA	Food and Drug Administration
FSP	Fine Structure Processing
HSSE	Harmonic Single Side-band Encoder
IBM	Ideal Binary Masking
IRM	Ideal Ratio Masking
MSE	Mean Square Error
NR	Noise Reduction
PC	Proportion Correct
PESQ	Perceptual Evaluation Score
PSOLA	Pitch Synchronous Overlap Add Algorithm
SDR	Signal to Distortion Ratio
SPEAK	Spectral Peak
SRTN	Speech Recognition Threshold In Noise
SS	Spectral Subtraction
SSE	Single Side-band Encoder
STOI	Short Time Objective Intelligibility
TFS	Temporal Fine Structures
TLE	Temporal Limit Encoder
WF	Wiener Filtering

Chapter 1

INTRODUCTION

The human ear is a complex organ that allows us to perceive sound. Sound waves travel through the outer ear and auditory canal, causing vibrations in the middle and inner ear. The cochlear fluid moves in the basilar membrane due to the pressure variations in the oval window, which causes the basilar membrane to vibrate. The tonotopic organization of the basilar membrane means that each location along it responds best to a particular frequency, which accounts for the spectral resolution of the human ear. The inner hair cells amplify and compress these vibrations, providing level and frequency-dependent gain control, enhancing the ear's sensitivity and frequency-resolving capabilities. Hair cells are fragile and easily damaged, resulting in hearing impairment or profound deafness. In such cases, cochlear implants can electrically stimulate the inner hair cells to transmit signals to the brain. Cochlear implants use a series of electrodes surgically placed into the cochlea of the inner ear. These electrodes electrically activate the auditory nerve fibers by bypassing the damaged auditory system, transmitting impulses to the brain. In cochlear implants, the speech processor plays a significant role in the optimal extraction and delivery of information from the input speech signal. The functional block diagram of commercially available cochlear implant (CI) is shown in Fig. 1.1. The human body's cochlear implantation is demonstrated in Fig. 1.2.

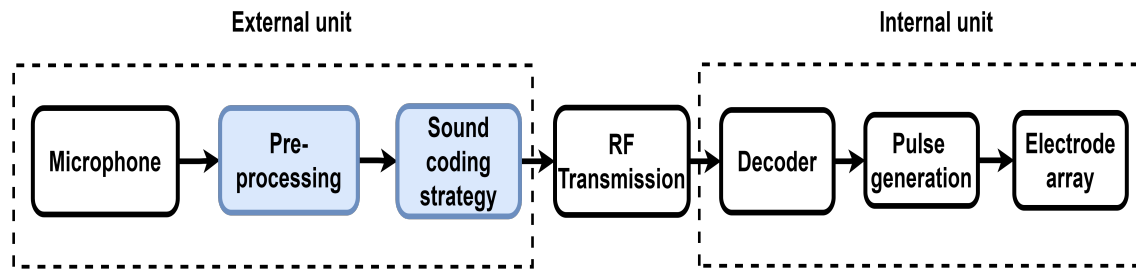


Figure 1.1: Functional block diagram of a cochlear implant.

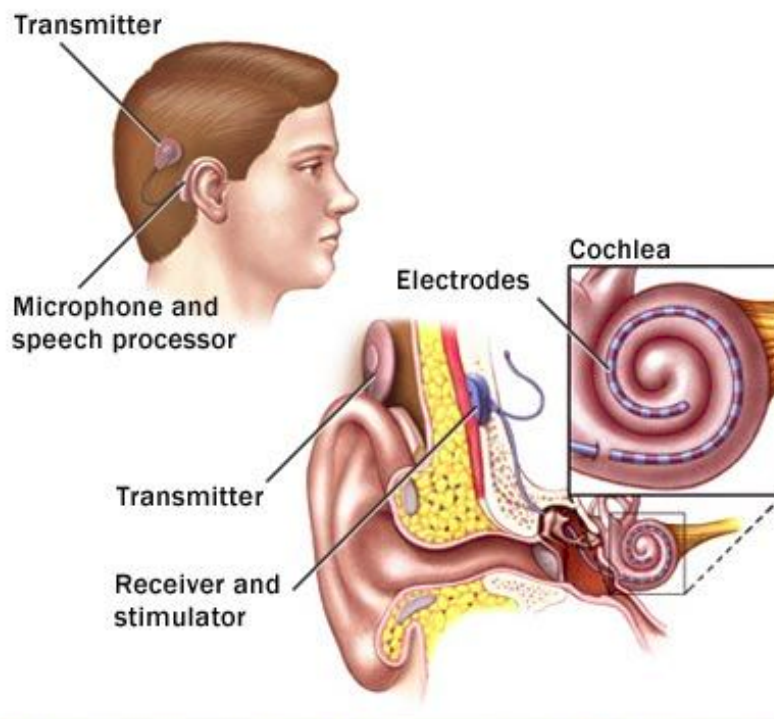


Figure 1.2: Block diagram of the cochlear implantation

Source: Mayo clinic

1.1 Motivation

The motivation behind this research stems from the fundamental goal of improving speech perception and intelligibility in individuals with cochlear implants. The outcomes of this research have far-reaching implications for individuals with cochlear implants, as improved speech perception and intelligibility would greatly enhance their communication abilities and overall quality of life.

Acoustic waveforms that are complex can be broken down into two components, a slowly changing envelope (ENV) and a fast-changing temporal fine-structure (TFS). One crucial aspect of speech perception is the ability to process TFS, which precisely encodes rapid changes in sound waveforms over time. TFS plays a crucial role in conveying speech nuances, such as pitch, prosody, and spectral details, all contributing to understanding speech in challenging acoustic conditions. However, current cochlear implant algorithms cannot faithfully represent TFS information, reducing speech intelligibility, especially in noisy environments.

Most of the sound coding strategies of CIs encode the envelope cue and discard the TFS (Moon and Hong 2014). The envelope cue is adequate for speech recognition in quiet, but are insufficient for speech recognition in background noise. This is because the temporal envelope does not effectively convey the essential cue for speech recognition in noise, such as fundamental frequency (F_0) and inter-aural time difference (Wouters *et al.* (2015)). Acoustic modeling studies have mentioned that providing an extra cue such as TFS would enhance speech recognition in noise (Nambi *et al.* (2016)). Among the various hypotheses that evolved over the years to explain the role of TFS in speech recognition in noise, a prominent one is TFS-mediated auditory stream segregation (Nie *et al.* (2004))(Guo *et al.* 2020). According to this hypothesis, TFS helps in segregating target speech and noise into two different streams and thus helps in the perception of target speech (Lorenzi *et al.* (2006))(Teng *et al.* 2019). Essential cues for stream segregation, such as (F_0) and harmonics, are weakly coded through ENV. However, the TFS can effectively carry the information of (F_0) and harmonics (Micheyl and Oxenham (2010)), and it is hypothesized that the ability of TFS to carry (F_0) and harmonics is the reason for better speech understanding in noise when the TFS is coded. Hence, developing a sound coding strategy to code TFS along with an envelope would improve the speech recognition ability of cochlear implantees in noisy environment.

In the field of cochlear implants, where auditory information is artificially delivered to the auditory nerve, the quality of the delivered signal is crucial for the overall performance of the device. However, the presence of background noise can degrade the quality of the auditory signal, compromising the user’s ability to accurately perceive and understand speech. CI user’s speech recognition scores are reduced from 60 to 30% when the signal-to-noise ratio is low in real-life situations (Spahr *et al.* (2007)). However, individuals with cochlear implants require a 25dB higher SNR to recognize

at minimum 50% of the target speech given in the background talker noise (Hast *et al.* 2015). Therefore, the development of noise reduction algorithms tailored specifically for cochlear implants is essential to address this issue and improve speech perception in noisy environments. These findings indicate that noise reduction strategies in CIs are a critical link in the signal processing pipeline as they help users maintain good speech intelligibility even in noisy conditions.

As a part of noise reduction/pre-processing for cochlear implants, deep denoising methods can be designed to adaptively estimate and reduce noise levels based on the input signal characteristics. Reducing noise and enhancing the quality of auditory signals can contribute to better speech perception, increased communication abilities, and improved overall hearing outcomes for individuals with cochlear implants. This adaptability can be especially useful in cochlear implants, where noise levels vary across different listening environments. Deep learning models typically require large amounts of clean data to perform well. However, in the case of low-resource languages, the lack of data poses a challenge. To avoid high dependence on the clean dataset to train the neural network, there is a need to design the training method with minimum reliance on a clean dataset.

1.2 Objectives of the Thesis

- **Objective 1:** Encoding frequency compressed temporal fine structure cues to improve speech recognition in noise in cochlear implants
- **Objective 2:** To study and implement an efficient noise reduction method for improving speech recognition in cochlear implants.
- **Objective 3:** To study and implement an efficient deep denoising method for improving speech recognition with minimal dependence on clean speech data.

1.3 Thesis Contributions

The research work presented in this study focuses on significantly enhancing speech intelligibility in cochlear implants. The contributions of this research encompass the development of innovative strategies to encode pitch-shifted temporal fine structures, the proposal of a cochlear implant-specific preprocessing method, and the introduction

of a novel deep denoising technique to improve speech perception in noisy conditions. The key contributions of this research are summarized as follows:

- Investigation of the significance of temporal fine structures (TFS) in speech perception in the presence of noise: The research examines the importance of TFS for robust speech understanding in noisy environments. By encoding various TFS cut-off frequencies, the study sheds light on the role of TFS information in enhancing speech intelligibility. Furthermore, a pitch-shifted method is proposed to encode TFS within the neuro-physiological limitations of cochlear implant users, taking into account their specific auditory processing capabilities.
- Proposal of a cochlear implant-specific preprocessing method for improved speech intelligibility: A novel preprocessing method is introduced to address the challenges associated with cochlear implants, such as limited perception of temporal fine structures. This method focuses on minimizing the mean square error between the enhanced speech and clean speech envelopes, thereby enhancing the accuracy of the auditory input. By maximizing speech understanding for cochlear implant users, this preprocessing technique aims to overcome the limitations imposed by the implant and optimize speech perception.
- Introduction of a novel deep denoising method with reduced dependency on clean speech data: This research presents a pioneering deep denoising method that overcomes the challenge of relying heavily on clean speech data for training deep learning models. In this approach, the deep learning model is trained exclusively using noisy speech data. By eliminating the requirement for extensive clean speech data, this method offers a practical solution to the scarcity of labeled clean speech datasets. The proposed deep denoising technique demonstrates promising results and shows potential for enhancing speech perception in real-world noisy conditions.

These contributions collectively advance the field of cochlear implants and pave the way for significant improvements in speech intelligibility for cochlear implant users. The proposed strategies for encoding TFS, the cochlear implant-specific preprocessing method, and the innovative deep denoising technique offer valuable insights and novel approaches for addressing the challenges faced by individuals with cochlear implants, ultimately contributing to the enhancement of their auditory experiences and overall quality of life.

1.4 Thesis Organization

The rest of the thesis is organized as follows.

- Chapter 2, presents an in-depth analysis of the sound coding strategies utilized in cochlear implants. The focus is on TFS coding strategies and the noise reduction techniques implemented to enhance speech intelligibility in cochlear implants.
- Chapter 3 explains the significance of TFS for CI users' speech perception in noise, along with creating an acoustic model that utilizes PSOLA to encode TFS while considering the neuropsychological constraints of CI users.
- Chapter 4 discusses an enhanced noise reduction technique that improves speech intelligibility in cochlear implants.
- Chapter 5 discusses the impact of Noisy2Noisy training on speech enhancement in DCU-net. The proposed method was compared to N2C and N2N methods using various metrics to measure intelligibility.
- Chapter 6 provides concluding remarks and outlines future research directions.

Chapter 2

BACKGROUND

Unfortunately, hearing loss is a widespread health issue, affecting over 360 million people globally and leading to hearing loss. The main cause of sensorineural hearing loss is the damage in the inner ear cochlea or the hair cells responsible for transmitting sound signals to the brain. This hearing loss is usually permanent and can be caused by various factors such as illness, exposure to loud noise, aging, use of certain drugs, and genetics. Fortunately, the Cochlear Implant (CI) is the most effective neural prosthesis for restoring hearing function. Medical prosthetic devices called cochlear implants consist of an electrode array surgically implanted into the cochlea. These devices provide electrical stimulation to the active auditory neurons. For individuals who experience moderate to profound hearing loss in one or both ears, as well as those who do not receive adequate benefits from hearing aids, cochlear implants can serve as a helpful solution.

2.1 Human auditory system

To perceive sound, the human auditory system detects vibrations through the ear. This system comprises sensory organs and auditory peripherals that work in tandem to process sensory information and enable our ability to hear. The human ear can detect sounds with great intensity and frequency resolution. The auditory system's frequency range falls between 20 to 20 kHz, with a normal conversation range of 300 to 3 kHz, and the dynamic range of human hearing is approximately 120 dB. The human auditory periphery comprises four sensory parts: the outer ear, middle ear, inner ear, and auditory nerves, as illustrated in Fig. 2.1.

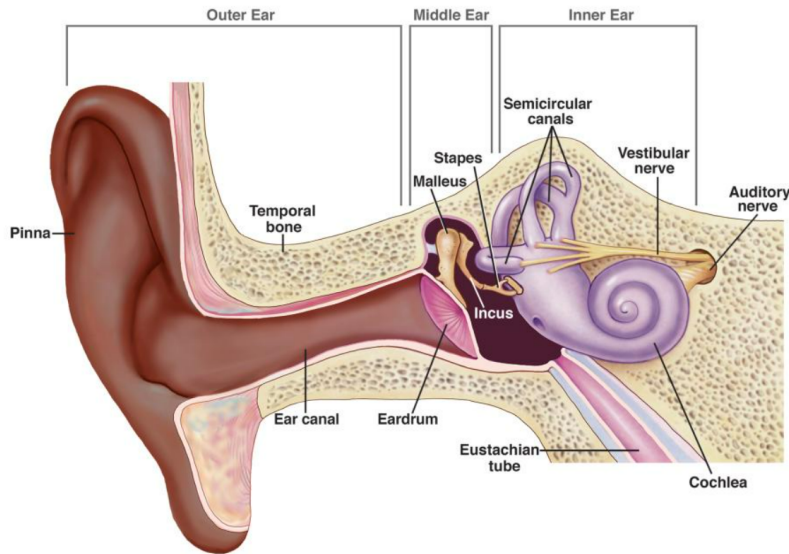


Figure 2.1: Schematic diagram of the human auditory system

Source: NIH/NIDCD

Hearing involves sound waves traveling through the outer ear, and the middle ear transfers the mechanical energy to the inner ear, where the vestibular nerves are stimulated. The outer ear comprises the auricle (pinna) and ear canal, while the middle ear is composed of the tympanic membrane (eardrum) and three ossicles (malleus, incus, and stapes). These bones are arranged in a way that allows for the amplification of sound waves and reduction of sound reflection. Conversely, the inner ear contains the cochlea - a spiral-shaped component consisting of three fluid-filled areas that are responsible for hearing sensation. As sound waves propagate from the middle ear towards the cochlea, they cause extracellular fluid in the cochlea to move along the basilar membrane, ultimately resulting in the sensation of hearing. The motion of the liquid in the membrane is detected by the hair cells with the help of stereocilia. This process converts sound waves' mechanical energy into electrical signals, stimulating the auditory nerves, resulting in excitation. The primary auditory neurons convert sound signals into electrochemical impulses, also known as action potentials. These impulses travel through the auditory nerve to brainstem structures for further processing. Acting as a filter, the basilar membrane transmits certain parts of sound waves to the auditory nerve fibers. The cochlea seems like a collection of parallel filters, known as auditory filters, with overlapping frequency ranges, to

conduct spectral analysis.

2.1.1 Human auditory filters

The cochlea contains a system of auditory filters known as the basilar membrane, which separates incoming sounds into different frequency components. This membrane is narrower and stiffer at the base (closest to the middle ear) and wider and more flexible at the apex (farthest from the middle ear). As a result, different parts of the basilar membrane vibrate maximally in response to different sound frequencies. The varying response characteristics of different regions along the basilar membrane correspond to different auditory filters. These filters are often referred to as "auditory filter bank." Each filter has a specific bandwidth and center frequency. The bandwidth of a filter refers to the range of frequencies it encompasses, while the center frequency represents the frequency at which it responds most strongly. Fig. 2.2 represents the sample responses of human auditory filters.

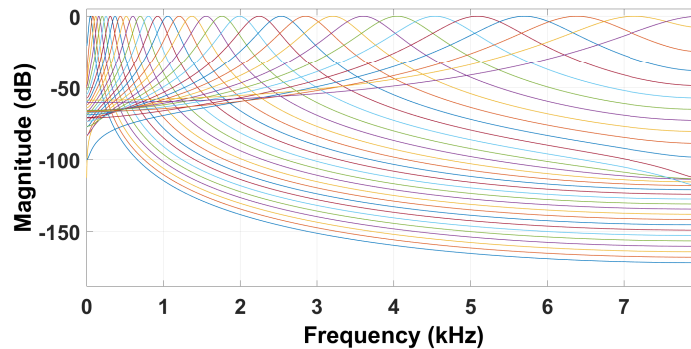


Figure 2.2: Responses of a sample of human auditory filters

2.2 Electric and Acoustic Stimulation

Electric stimulation is the primary method used in cochlear implants. The external processor captures sounds from the environment, processes them into electrical signals, and sends them to the internal component. The internal electrode array is surgically inserted into the cochlea, where it stimulates the auditory nerve fibers. The electrode array consists of multiple electrodes that are strategically placed along the cochlea. Each electrode corresponds to a specific frequency range of sound. When the electrical

signals reach the electrodes, they stimulate the surrounding auditory nerve fibers, bypassing the damaged hair cells in the cochlea. The auditory nerve fibers then carry these signals to the brain, where they are interpreted as sound.

In recent years, there have been advancements in cochlear implant technology that incorporate both electric and acoustic stimulation. These devices are known as hybrid or electro-acoustic cochlear implants. Hybrid cochlear implants combine the electrical stimulation the internal electrode array provides with the acoustic amplification of low-frequency sounds. These devices are designed for individuals who have residual low-frequency hearing. The low-frequency sounds, which are not effectively captured by the electrode array, are amplified and delivered through a traditional hearing aid device. The high-frequency sounds are still processed and delivered through electrical stimulation.

2.3 Commercially available cochlear implants

The criteria for determining cochlear implant candidacy depend on the patient’s medical status. The FDA has established guidelines based on clinical investigations to ensure the safety and effectiveness of the implants. Currently, three manufacturers have received FDA approval for producing cochlear implant processors. These include Cochlear Corporation’s Nucleus processors, which use the spectral peak (SPEAK) strategy; Advanced Bionics Corporation’s Clarion devices, which use either compressed analog (CA) or continuous interleaved sampling (CIS) strategies; and Medical Electronics Corporation’s processors, which utilize high-rate CIS or high-rate SPEAK strategies. These processors have significantly improved the functionality of cochlear implant devices and have played a crucial role in enhancing hearing for hearing-impaired individuals over the past decade. The signal-processing methods used in the processors mentioned above are discussed in the next section. Table 2.1 shows some coding strategies that are utilized in commercially available cochlear implants.

Table 2.1: A comparison of different coding strategies used in commercial CIs.

Method	No.electrodes	Allows TFS	Used in	Benefits	Remarks
CIS	16	NO	Clarion processor and Nucleus CI24M device (Cochlear)	No channel interaction	Poor performance in noise
SPEAK	20	NO	Nucleus Spectra 22 (Cochlear)	Stimuli delivered to selected electrodes	The pulse rate of the stimuli varies
ACE	20	NO	Nucleus 24 (Cochlear)	The stimuli can be delivered in two ways	There is no guarantee that the bands selected by this approach are those with peak amplitudes.
FSP	12	YES	MED-EL	Speech perception to music enjoyment	Speech intelligibility is limited
HiRes120	16	YES	Advanced Bionics	Hi resolution:delivers more (120) spectral bands	No clear significant improvement in speech and music perception

2.4 Signal Processing Strategies In Cochlear Implants

Signal processing strategies in cochlear implants play a crucial role in converting sound into electrical signals that can be interpreted by the auditory nerve and perceived as sound by the recipient. Here are some common signal-processing steps employed in cochlear implants:

- **Sound Processing:** The cochlear implant system consists of an external sound processor and an internal implant. The sound processor captures sounds from the environment using a microphone and then processes them. The sound processing algorithms aim to enhance the important acoustic cues for speech, such

as fundamental frequency (pitch), temporal envelope (amplitude modulation), and spectral information. Various techniques like noise reduction, dynamic range compression, and directional microphones are employed to improve the signal quality.

- **Feature Extraction:** After sound processing, the signal is converted into electrical stimulation patterns that can be delivered to the auditory nerve. Feature extraction algorithms analyze the processed sound and extract relevant acoustic features, such as spectral peaks or modulation rates. These features provide information about different aspects of the sound, which is then used to determine the appropriate electrical stimulation patterns.
- **Electrical Stimulation:** Once the features are extracted, they are mapped onto electrical stimulation patterns. The cochlear implant's internal electrode array is placed within the cochlea, and each electrode stimulates a specific region along the tonotopic gradient. The electrical stimulation patterns are designed to mimic the frequency and temporal characteristics of the original sound signal. The intensity, timing, and electrode selection are optimized to give the user the best possible sound perception.

Signal processing strategies for cochlear implants are continually evolving, and several different approaches have been developed over the years. Some common strategies include:

- **Continuous Interleaved Sampling (CIS):** This strategy uses a high-rate pulse train with a constant stimulation rate across electrodes. It provides temporal information but sacrifices spectral resolution.
- **Spectral Peak (SPEAK):** SPEAK strategy focuses on preserving the spectral information by using amplitude-modulated pulse trains and prioritizing the stimulation of spectral peaks.
- **Advanced Combination Encoder (ACE):** ACE strategy employs simultaneous stimulation across electrodes and incorporates information about both the spectral and temporal aspects of sound.
- **Fine Structure Processing (FSP):** FSP aims to convey the fine temporal structure of sound by utilizing the fine-structure cues. It employs a high pulse rate and

emphasizes the timing of individual pulses.

These strategies, along with their variants and combinations, are designed to optimize speech perception and sound quality for cochlear implant users. Researchers and engineers continuously work to improve signal processing algorithms to enhance the performance and usability of cochlear implant devices.

2.4.1 Continuous Interleaved Sampling

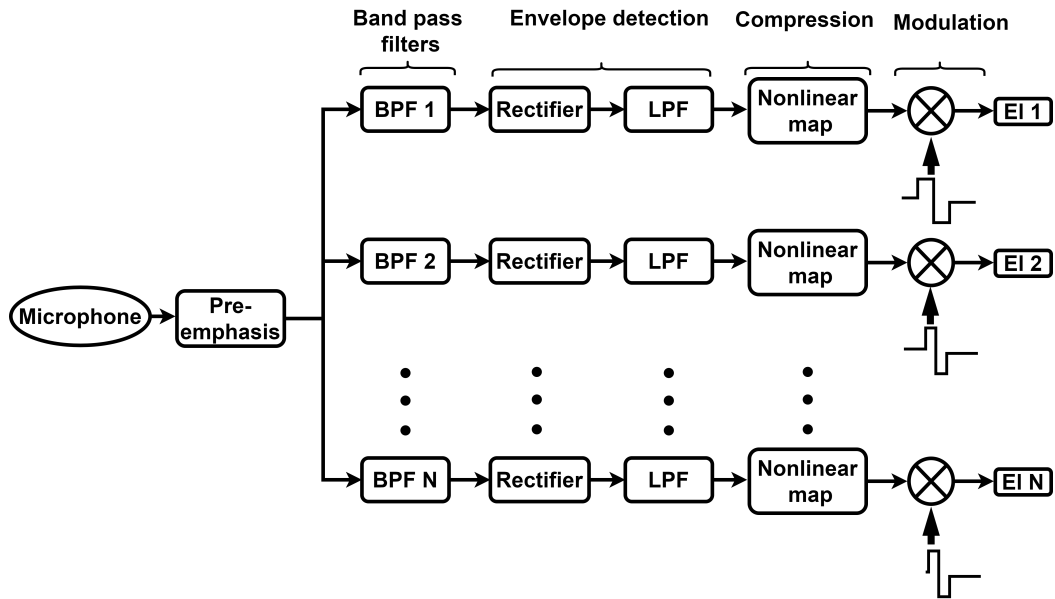


Figure 2.3: Block diagram of the CIS stimulation strategy

Continuous Interleaved Sampling (CIS) is a well-known and commonly used technique in cochlear implants for speech processing. This method utilizes the Pulsatile Waveform technique and follows the CIS algorithm flow, as depicted in Fig. 2.3. The input signal is first subjected to pre-emphasis and then filtered using a filter bank consisting of n bandpass filters with non-linear bandwidths. The preferred value of n is typically eight in an n -channel CIS algorithm. The signals then pass through the filter bank, and their envelopes are computed through full-wave rectification and a low-pass filter with a standard cut-off of 50 Hz. During conversations, volume levels can fluctuate up to 30 dB, but individuals with hearing implants may have a smaller range of only 5 dB. To compress envelopes, acoustic amplitudes are mapped to electrical amplitudes using non-linear mapping functions. These mapping functions help to map acoustic amplitudes to the patient’s electrical dynamic range.

2.4.2 Spectral peak (SPEAK) Strategy

The Spectral peak (SPEAK) strategy is a unique approach to spectral analysis compared to other strategies. This method utilizes a 20-channel bandpass filter bank to filter input speech signals. The amplitude detection module is then used to detect channel amplitudes. Spectral maxima are obtained for each channel amplitude by comparing them to a base value. To stimulate the appropriate electrodes in a tonotopic order, only channel amplitudes that exceed the base value are used. Stimulating only the electrodes corresponding to the spectral maxima, the process starts from the base and moves to the apex. The frequency of the stimuli differs depending on the number of electrodes stimulated in each cycle. The Nucleus Spectra 22 processor employs this signal-processing strategy.

2.4.3 Advanced Combination Encoder (ACE)

The Nucleus implant utilizes the advanced combination encoder (ACE) strategy, which falls under the waveform representation and is classified as an "N of M" type strategy. The spectral peak (SPEAK) strategy shares many similarities with the ACE strategy but differs in its pulse per second (PPS) rate. Figure 2.4 presents a block diagram that fundamentally represents the ACE strategy.

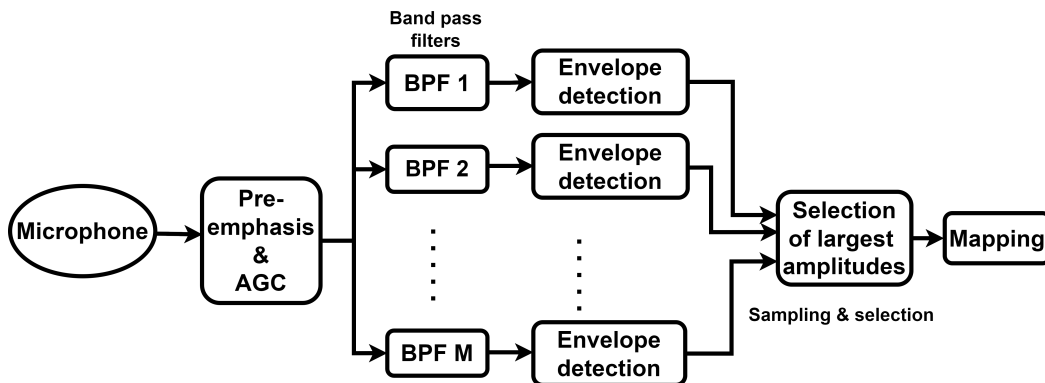


Figure 2.4: Block diagram illustrating ACE.

A filter is used to boost the high-frequency components to enhance the signal from the microphone. An adaptive gain control (AGC) reduces amplification at the appropriate time to prevent distortion of loud sounds. Once the signal is captured, it undergoes digitization and is transmitted through a filter bank. The filter bank

consists of linearly spaced frequency bands below 1000 Hz and logarithmically spaced bands above 1000 Hz.

To extract the envelope from the audio signal, each spectral band is analyzed by computing the magnitude of the complex output. Each electrode is assigned to a bandpass filter, which represents a channel. For every audio signal frame, a stimulation cycle is completed by sequentially stimulating N electrodes. The stimulation rate on a single channel is determined by the number of cycles per second, also known as the channel stimulation rate.

The number of channels (electrodes) and overall stimulation rate limit the bandwidth of a cochlear implant. The implant's temporal resolution is represented by the channel stimulation rate while the frequency resolution is represented by the total number of electrodes, M . However, only a subset of filter bank output samples with the largest amplitude is selected as N out of M electrodes ($N < M$) are stimulated in each cycle. Reducing N improves the channel stimulation rate, providing a better temporal representation of the audio signal but deteriorates the spectral representation of the audio signal. Conversely, if the channel stimulation rate is decreased, N can be improved, providing a better spectral representation of the audio signal. Finally, the last stage of the process compresses the acoustic amplitudes into the subject's dynamic range between the measured threshold and maximum comfortable loudness level for electrical stimulation by mapping the amplitudes to the corresponding electrodes.

2.4.4 Fine Structure Processing (FSP)

People who use cochlear implants experience effective sound processing in quiet environments through methods like ACE and CIS. However, speech intelligibility can be inadequate in noisy environments, and specific aspects of music, like pitch, may not be perceived correctly. One reason could be the absence of temporal fine structure (TFS) in the stimulation patterns. TFS refers to the rapid fluctuations in sound waves that provide information about pitch, timbre, and location. The goal of FSP is to maintain and convey these fine details by using the natural encoding mechanisms of the auditory system.

The functional block diagram of the FSP strategy is shown in Fig. 2.5. The purpose of FSP is to accurately represent TFS data within the lowest frequencies of the input sound signals. This is achieved by transmitting bursts of stimulus beats through one or a few corresponding CI electrodes. These bursts may consist of single

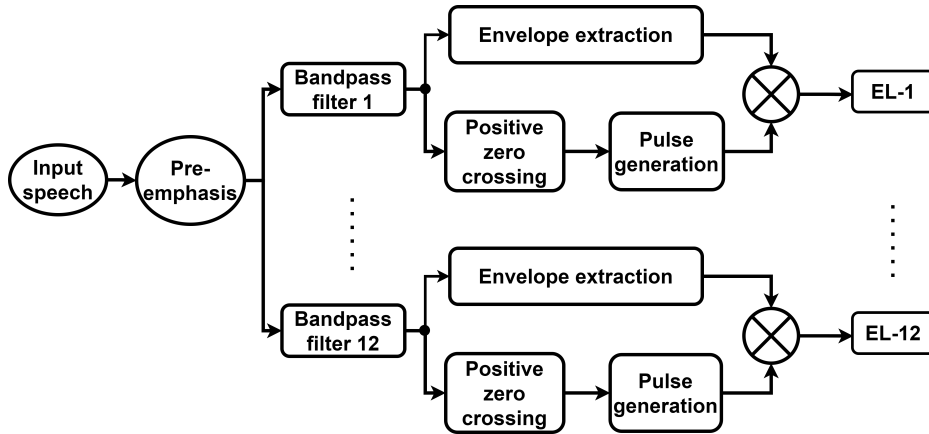


Figure 2.5: Block diagram of FSP strategy

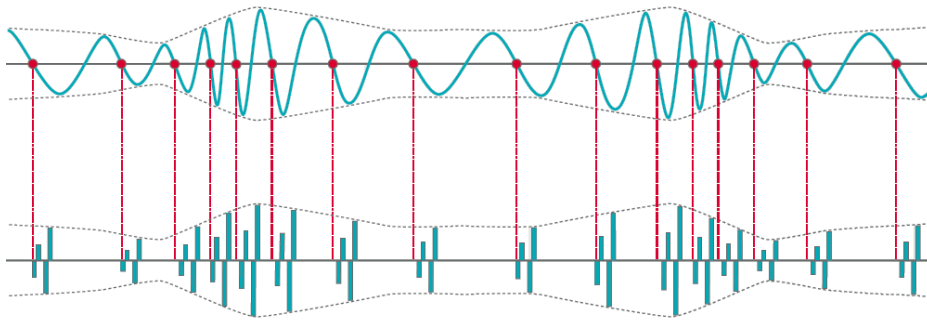


Figure 2.6: Generating pulses based on positive zero crossings
Source: MED-EL

or multiple stimulation beats and are determined indirectly through band-limited acoustic signals. A positive zero-crossing triggers the bursts in the band pass-filtered waveform, as shown in Fig. 2.6.

Each burst's length and amplitude-envelope modulation is predetermined to approximate the filtered acoustic waveform after half-wave rectification. These bursts contain valuable data about the TFS in the lower frequency bands, which cannot be accessed through the envelope of the signals. This may lead to improved perception for CI users. FSP utilizes variable-rate coding to offer additional TFS data. Med-El offers different versions of FSP, such as FS4 and FS4-p coding strategies. These strategies differ primarily in the frequency range over which TFS is presented. FSP represents TFS for frequencies up to 350-500 Hz, while FS4 and FS4-p provide TFS

for frequencies up to 750-950 Hz. By default, these techniques cover a frequency range of 100-8500 Hz to represent F_0 accurately. It is worth noting that this range differs from the CIS strategies from Med-El, which cover 250-8500 Hz.

2.4.5 HiRes120

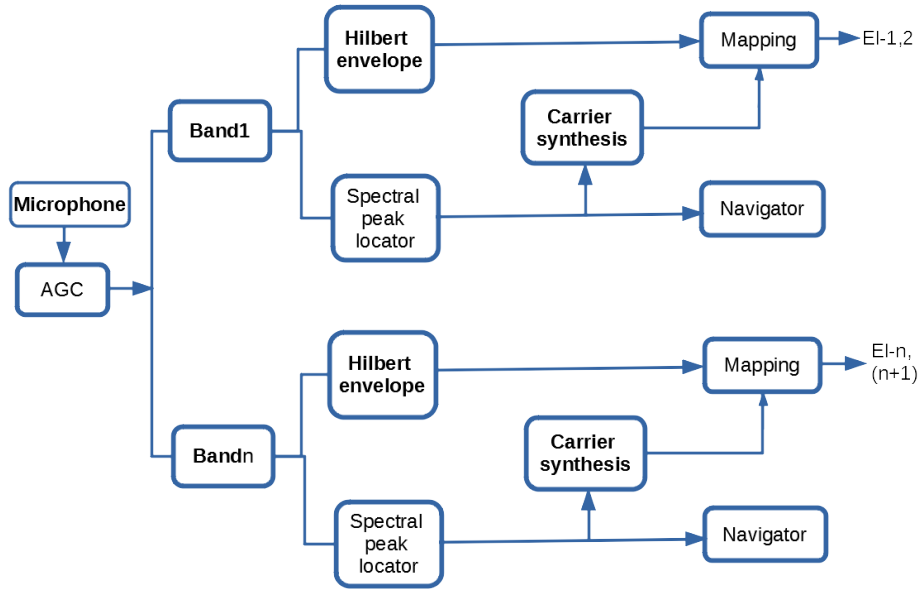


Figure 2.7: Block diagram of HiRes120 strategy

In Advanced Bionics systems, a sound processing technique called HiRes120 aims to improve the delivery of TFS information to CI recipients. The functional block diagram of the HiRes120 is shown in Fig. 2.7. This technique involves identifying the dominant spectral peak within each band-pass filter used for spectral analysis of incoming sounds. Using the frequency of each spectral peak to control a synthetic modulator, modulations containing temporal information are added to each frequency band. These modulations are combined with the corresponding envelope levels and then sampled in synchrony with the pulses delivered to the electrodes. The estimated peak frequency within each analysis filter is also used to control the relative currents of pulses delivered simultaneously on two adjacent electrodes allocated to the filter. Virtual channels can be formed by adjusting the relative currents on the electrode pairs. This enables more precise spatial resolution in directing the location of maximum neural activity compared to activating the electrodes individually. It should be

noted that the Advanced Bionics implant has 16 intracochlear electrodes, allowing for the allocation of 15 paired electrodes to the filters.

Psychophysical considerations have detailed proof that most CI recipients cannot resolve temporal patterns from adjacent electrodes. This proposes that sound-processing techniques like HiRes120 and FSP, which use exceptionally different approaches but depend on giving independent information channels across adjacent electrodes, may result in limited benefits (McKay and McDermott 1996). Further studies should be conducted on the listening experiences of CI recipients using schemes like HiRes120 and FSP over a longer period of time to determine if these schemes provide perceptually beneficial fine structure information.

Several coding strategies were proposed to encode more harmonics through TFS by frequency shifting, such as the Single Side-band Encoder (SSE) (Nie *et al.* 2008), Harmonic Single Side-band Encoder (HSSE) (Li *et al.* 2010), and Temporal Limits Encoder(TLE) (Meng *et al.* 2015). Detailed explanations of these coding strategies are provided in upcoming sections.

2.4.6 Single Side-band Encoder (SSE)

The ability of cochlear implants to encode temporal fine structures is limited by the patient’s ability to perceive electrical stimulation. Research, including Zeng’s study, has shown that cochlear implant patients can only perceive variations in stimulation rate up to 1000 Hz. However, the frequency content of temporal fine structure in speech and music can be up to 10,000 Hz at higher sub-bands, making it a non-band-limited signal. To overcome this limitation, the SSE method uses a single sideband demodulation approach to shift a sub-band signal to its baseband and generate a low-frequency, coherent envelope signal that carries both temporal envelope and fine structure cues in a slowly-varying manner. This method makes it possible to deliver perceivable temporal cues to cochlear implants. Through an acoustic simulation experiment, SSE performed better than CIS in recognizing melodies.

2.4.7 Harmonic Single Side-band Encoder (HSSE)

Li *et al.* (Li *et al.* 2010) introduced a new method to enhance SSE called ”harmonic coherent demodulation” (HSSE). Unlike SSE, which defines the carrier frequency of the sub-band signal as the lower cut-off frequency of the respective sub-band, HSSE se-

lects an integer multiple of the fundamental frequency (F0), also known as a harmonic frequency, as the carrier frequency. This allows for the coherent envelope to fluctuate in sync with the instantaneous F0, resulting in improved pitch-related perception and speech perception in noisy environments. Experimental studies have shown that HSSE outperforms CIS (Li *et al.* 2013).

2.4.8 Temporal Limits Encoder (TLE)

The temporal limit encoder (TLE) (Meng *et al.* 2015) method suggests a carrier frequency of 50 Hz and an even distribution of bandwidth, with each channel having a narrow bandwidth. This setup allows for different frequencies within the continuous speech frequency range (typically 50-3250 Hz for a 16-channel scenario) to be expressed as varying amplitude periodicities in specific channels. This characteristic, based on psychophysical knowledge, can be ranked on a pitch scale. Other strategies like CIS, SSE, and HSSE do not have this feature. The TLE strategy is an extension of SSE and has been developed to enhance CI performance.

2.5 Noise Reduction Methods for cochlear implants

Background noise can make it difficult for cochlear implant users to understand speech, particularly in noisy environments such as crowded rooms, restaurants, or public spaces. Noise reduction methods aim to suppress or minimize background noise, thereby improving speech perception and making it easier for individuals to communicate effectively. The signal-to-noise ratio (SNR) is an essential factor in auditory perception. It represents the desired sound level compared to the background noise level. A higher SNR indicates a more favourable listening environment. Noise reduction methods aim to increase the SNR by reducing background noise, allowing the user to focus on the desired sound and improving their ability to understand speech and other auditory cues.

Over the years, various techniques have been proposed to reduce noise and improve speech quality in the presence of background noise (Henry *et al.* 2021). These algorithms typically rely on either assumptions or preprocessing methods. Time-frequency masking methods are commonly used in cochlear implants to improve speech perception. The following time-frequency masking methods are discussed below.

2.5.1 Ideal Binary Mask

The ideal binary mask (IBM) is computed by comparing the magnitude spectrogram of the target signal with the estimated noise power spectral density. For each time-frequency bin, if the magnitude of the target signal is greater than the noise estimate, the binary mask is set to 1 (indicating presence of speech), otherwise it is set to 0 (indicating absence of speech). The Ideal binary mask is defined as follows

$$IBM(\tau_s, \omega) = \begin{cases} 1 & \text{if } \gamma(\tau_s, \omega) > \gamma_{in} \\ 0 & \text{if } \gamma(\tau_s, \omega) \leq \gamma_{in} \end{cases} \quad (2.1)$$

Where $\gamma(\tau_s, \omega)$ is short-term SNR and γ_{in} is overall input noisy speech SNR.

2.5.2 Wiener Filter

The Wiener filter (WF) can help to reduce this background noise by estimating the clean speech signal from the noisy input. The binary mask has values of either 0 or 1, indicating the absence or presence of a source in each time-frequency bin. This approach works well when the sources are well-separated and non-overlapping. However, in cases where the sources overlap in time and frequency, binary masks can cause artifacts and distortions in the separated signals. Soft time-frequency masking or WF method address this issue by using continuous-valued masks that smoothly transition between 0 and 1. This allows for a more flexible and gradual separation of the sources, resulting in improved sound quality. The WF mask is defined as follows:

$$W(\tau_s, \omega) = \frac{\gamma(\tau_s, \omega)}{1 + \gamma(\tau_s, \omega)} \quad (2.2)$$

2.5.3 Sigmoidal function

The sigmoid function is designed to heavily reduce the signal-to-noise ratio in channels with low SNR, while channels with high SNR experience little to no attenuation. The sigmoid function can be defined as follows:

$$g(\tau_s, \omega) = e^{-2/\gamma(\tau_s, \omega)} \quad (2.3)$$

Where $g(\tau_s, \omega)$ is a weighting function, its values vary between 1 and 0. The function's weight remains consistent at one when the SNR is greater than 20 dB but drops to zero when the SNR is less than -5 dB. The function above was selected because it resembles the sigmoidal shape of the human listener's psychometric function for intelligibility against SNR.

2.5.4 Spectral Subtraction

Spectral subtraction (SS) is used to improve the speech signal and minimize background noise. This process involves obtaining a noise estimate by analysing portions of the incoming signal that contain little or no speech information. The noise spectrum is subtracted from the noisy input signal to enhance the speech components. This is typically done in the frequency domain. After subtracting the noise spectrum, the resulting signal may be amplified to ensure appropriate loudness perception for the recipient. While spectral subtraction can reduce background noise and improve the signal-to-noise ratio for cochlear implant systems, it may not always produce optimal results in complex listening environments and can introduce some artifacts.

2.5.5 Ideal Ratio Mask

One commonly used technique for improving speech quality is Ideal Ratio Masking (IRM). The primary objective of this method is to isolate the desired speech signal from the surrounding noise. IRM works by identifying the time-frequency sections of an audio signal that contain the target speech and minimizing the impact of noise in those areas. The IRM is defined as

$$IRM(\tau_s, \omega) = \sqrt{\frac{X(\tau_s, \omega)}{N(\tau_s, \omega) + X(\tau_s, \omega)}} \quad (2.4)$$

The variable $X(\tau_s, \omega)$ represents the amount of speech energy in a given time-frequency unit (τ_s, ω) , while $N(\tau_s, \omega)$ represents the amount of noise energy in that same unit (τ_s, ω) . The Ideal Ratio Mask (IRM) is a method of scaling each T-F unit based on its signal-to-noise ratio (SNR), such that units with higher SNRs are attenuated less while those with lower SNRs are attenuated more. This differs from the Binary Mask (IBM) approach, which classifies T-F units as speech-dominant or noise-dominant based on a local SNR (γ_{in}) criterion. In IBM, speech-dominant units with SNR values greater

than γ_{in} are retained, while noise-dominant units with SNR values less than or equal to γ_{in} are discarded entirely.

Table 2.2 provides a detailed summary of time-frequency masking methods used as noise reduction techniques in cochlear implants to improve speech intelligibility.

Table 2.2: Summary of time frequency masking methods for noise reduction

Authors & Year	Mask	Noise types	Subjective test	Objective metric
Plapous et al. 2006	WF	Street, Car, and Babble	24 NH listeners	CD, SSNR
Hu et al. 2007	Sigmoidal	Babble	9 CI users	ANOVA(p<0.005)
Kim et al. 2009	IBM	Babble, Speech-shaped noise babble	17 NH Listeners	ANOVA(p<0.005)
Hazrati et al. 2013	IRM	Reverberant noise	6 CI users	ANOVA(p<0.005)
Koning et al. 2014	IBM	Speech/Babble	6 NH, and 6 CI users	ANOVA(p<0.005)
Chen et al. 2015	logMMSE, SS, WF	Babble noise	7 CI users	ANOVA(p<0.005)
Healy et al. 2015	IRM	Babble, Speech-shaped noise babble	10 NH, 7 CI users	STOI
Lai et al. 2016	MMSE	Cocktail, 2-talker babble	10 NH listeners	STOI, NCM
Purdy et al. 2017	SNR-NR	4T- babble noise	13 CI users	ANOVA(p<0.005)
Koning et al. 2018	IWF/IBM	Babble	6 NH,9 HI	ANOVA(p<0.005)
Wang and Hansen 2018	MMSE	Speech-shaped noise, Babble noise	6 CI user	ANOVA(p<0.005), WRR
Chiea et al. 2019	WF	Babble, white noise	-	STOI, PESQ, ANOVA(p<0.005)
Mourao et al. 2020	MMSE	Babble noise	4 NH & 6 CI users	STOI, NCM, SRMR-CI
Zhou et al. 2020	IRM	SSN and babble noise	11 CI users	ANOVA(p<0.005)
Chiea et al. 2021	WF	Artificial 4-talker babble	6 NH & 6 CI users	ANOVA(p<0.005), SRMR-CI

Chapter 3

DESIGN OF COCHLEAR ACOUSTIC MODEL TO ENCODE TEMPORAL FINE STRUCTURES

In this chapter, the significance of the TFS cut-off frequencies in cochlear implant (CI) speech coding for better speech perception in noise has been investigated. Based on the observations, an algorithm is proposed to represent TFS as proportionally frequency compressed cues. In order to encode TFS within the neuro-physiological limitations of the CI users, a pitch-shifted overlap-add algorithm (PSOLA) is proposed. The speech recognition scores (SRS) were measured at -10dB to +10dB for eight signal processing conditions corresponding to sinewave vocoder without TFS (NO-TFS), four unshifted TFS conditions including full band TFS, TFS up to 2000, 1000, and 600 Hz, and three conditions with PSOLA which shifted 2000, 1000 and 600 Hz TFS to 1000, 500 and 300 Hz respectively. The original envelope was unchanged across the conditions. Hence, the objective has been subdivided into two parts: First, to investigate the significance of temporal fine structures with different cut-off frequencies for improving speech intelligibility. The second one is encoding the TFS within the neuro-physiological limitations of the CI users.

3.1 Introduction

CIs primarily encode temporal envelope (ENV) and discard temporal fine structure (TFS) in individual channels (Moon and Hong 2014). Thus, lack of TFS is viewed as one of the reasons for poor speech recognition with CI in the presence of noise (Stronks *et al.* 2020, Tejani and Brown 2020, Dhanasingh and Hochmair 2021). Acoustic simulations of CI using sinewave vocoders have revealed that, addition of TFS along with ENV significantly improves speech recognition ability in noise (Meng *et al.* 2016). Among the various hypotheses evolved over years to explain the role of TFS in speech recognition in noise, a prominent one is TFS mediated auditory stream segregation (Nie *et al.* 2004, Teng *et al.* 2019). According to this hypothesis, TFS helps in segregating target speech and noise into two different streams and thus helping in the perception of target speech (Lorenzi *et al.* 2006, Moore 2021). Essential cues for stream segregation such as fundamental frequency (F_0) and harmonics are weakly coded through ENV (Dhanasingh and Hochmair 2021). However, the TFS can effectively carry the information of F_0 and harmonics (Micheyl and Oxenham 2010, Bianchi *et al.* 2019) and it is hypothesized that the ability of TFS to carry F_0 and harmonics is the reason for better speech understanding in noise when TFS is coded.

The perceptual benefit of TFS has motivated the manufacturers to implement the TFS encoding in the CI sound coding strategies (Dhanasingh and Hochmair 2021). One such commercially available sound coding strategy is channel specific sampling sequence (CSSS) (Hochmair *et al.* 2006). Based on CSSS, the MED-EL has proposed TFS coding strategies such as fine structure processing (FSP), FS4, and FS4-p (Wouters *et al.* 2015). The primary difference between these techniques is the frequency range in which TFS is delivered. While TFS frequencies in FSP are provided up to 350–500 Hz, FS4 and FS4-p provide TFS frequencies up to 750–950 Hz (Müller *et al.* 2020). In these strategies, TFS is encoded by varying the pulse rate to mimic the TFS variation. In CSSS, the TFS is estimated as the positive zero-crossings and is encoded as double burst of bi-phasic pulses corresponding to zero crossings (Riss *et al.* 2014). Therefore the CSSS strategy encodes both TFS and ENV in the low frequency channels whereas in the high frequency channels only the ENV is coded. Experimental studies comparing the variants of CSSS strategies (Eg:- FSP, FS4, FS4-p) and ENV-only strategies revealed mixed results ranging from no benefit to significant benefit (Fischer *et al.* 2021). The equivocal findings for the benefit of CSSS can be attributed

to the neuro-physiological limitations in the temporal coding (Müller *et al.* 2018). TFS is encoded in the auditory system by synchronizing nerve spikes to a specific phase of the signal, known as phase locking or temporal coding (Joris and Yin 1992, D’Alessandro *et al.* 2018, Guo *et al.* 2020).

Based on the above literature review, the research gap can be summarized as follows:

- Zeng 2002 reported that, in most CI recipients the temporal encoding saturates by 300 pulses per seconds (pps), and in some extremely well-performing candidates the temporal coding saturates by 1000pps. This temporal encoding saturation may limit the benefit of TFS offered by the coding strategy.
- One of the major advantage of TFS is the facilitation of pitch mediated stream segregation as it preserves F_0 and harmonicity. The strength of the complex signal’s pitch perception depends on the F_0 (Zeremadini *et al.* 2017) as well as harmonics. Nevertheless, the temporal coding saturation might limit the number of harmonics coded in the auditory system.
- A number of coding strategies were proposed to encode more harmonics through TFS by frequency shifting such as the Single Side-band Encoder (SSE) (Nie *et al.* 2008) , Harmonic Single Side-band Encoder (HSSE) (Li *et al.* 2010), and Temporal Limits Encoder(TLE) (Meng *et al.* 2015, Kan and Meng 2020). However none of these coding strategies effectively encode the harmonics within the temporal limits of CIs (Fischer *et al.* 2021).

The work investigates the significance of the required TFS cut-off frequencies in CIs. Hence, this work proposes a coding strategy, pitch synchronous overlap-add algorithm (PSOLA), involving the implementation of a proportional frequency compression to enable more harmonics to be coded within the temporal encoding limit of CI listeners for better perception in noise. The speech recognition score (SRS) was measured for the eight conditions that included sinewave vocoder (NO-TFS), full-band TFS, TFS up to 2000 Hz (lower 11 channels), TFS up to 1000 Hz (lower 8 channels), TFS up to 600 Hz (lower 6 channels), TFS 2000 Hz PSOLA, TFS 1000 Hz PSOLA, TFS 600 Hz PSOLA. In PSOLA, encoded TFS was downward shifted by a fixed pitch scaling factor of 0.5. In all the signal processing conditions, ENV was unchanged. The proportion correct (PC) responses were compared across the conditions from full band TFS to sinewave vocoder.

The following sections of the chapter are organized as follows: Section 3.2 investigates the significance of TFS for improving speech intelligibility in noise. Section 3.3 gives the detailed explanation of the pitch-shifted TFS. Section 3.4, discuss the simulation results of the proposed method. Section 3.5 of this work provides a discussion. Summary of the chapter is provided in section 3.6.

3.2 Investigating the significance of TFS for improving speech intelligibility

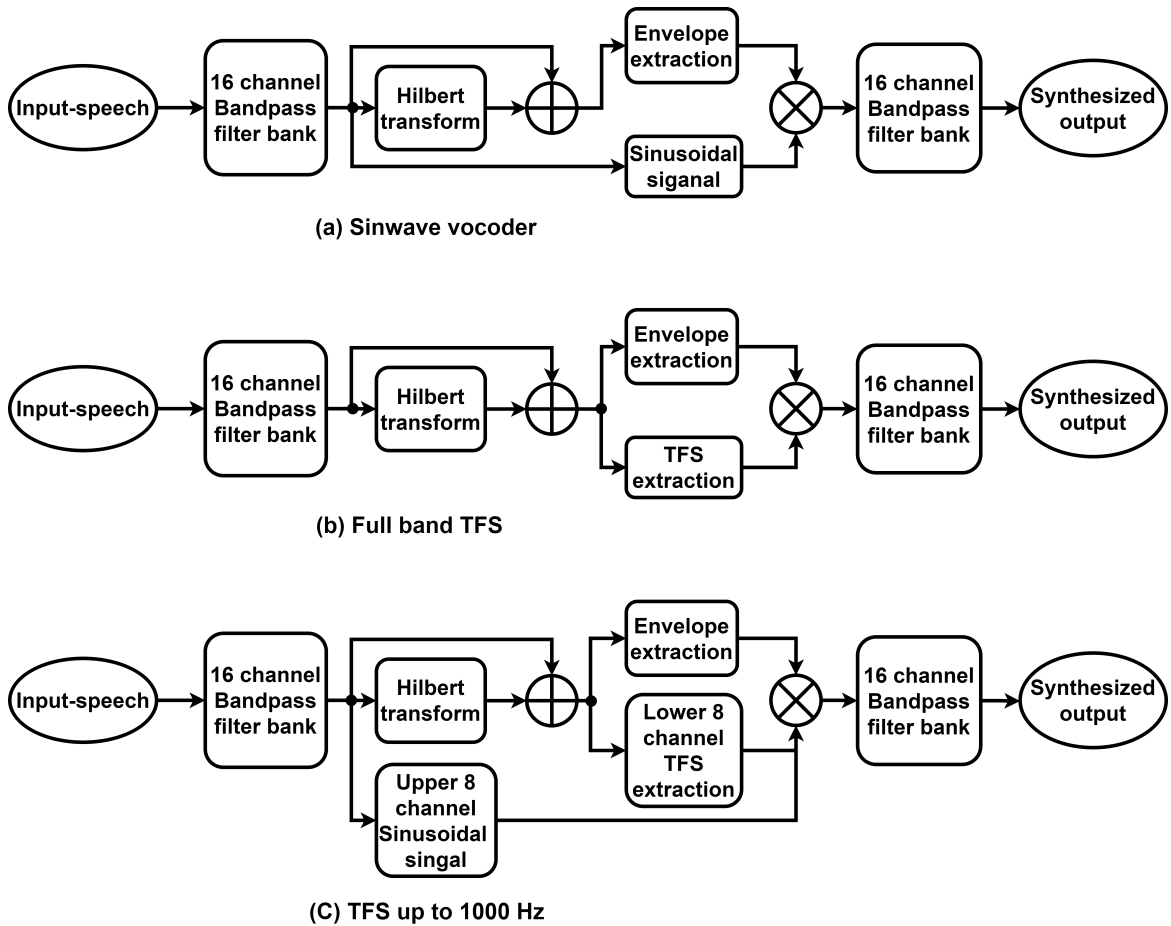


Figure 3.1: The block diagram of the three speech encoding methods in CIs

In this work an acoustic cochlear implant model was designed that can encode Full band TFS and TFS frequency up to 1000 Hz to investigate the significance of TFS for speech intelligibility in cochlear implants. The cut-off frequency of 1000 Hz was

selected for TFS based on some well performing CI users’ preferences (Müller *et al.* 2020, Zeng 2002). In order to understand speech intelligibility in cochlear implants, this study first took into account the commercially available acoustic cochlear implant model, sinewave vocoder, as illustrated in Fig. 3.1(a). The sinewave vocoder mainly consists of four steps: the first step is to design a filter bank that should be compatible with the human auditory filter bank. The second step is to extract the envelope and phase information using Hilbert transform. In the third step, the extracted envelope is modulated by a sinusoidal signal with the corresponding centre frequency (f_c) of the band pass filter. Finally, the modulated signal is passed through the synthesis filter bank for better reconstruction.

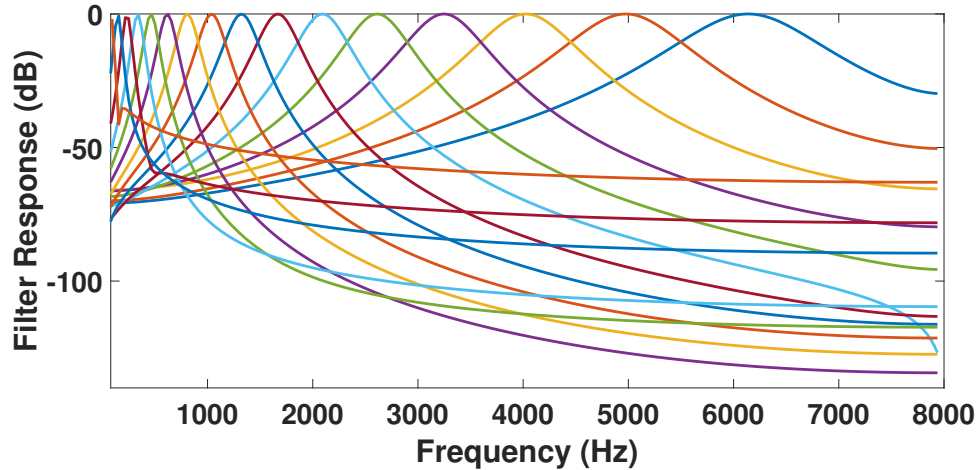


Figure 3.2: Filter bank response for 16 bands

First step that involved a filter bank imitates the human cochlea. The filter bank, which is constructed as a set of parallel bandpass filters each tuned to a distinct frequency, from the basis of Patterson’s cochlear model (Slaney *et al.* 1993). From high frequencies at the cochlea’s base to low frequencies at its apex, the filters are arranged tonotopically. An Equivalent Rectangular Bandwidth (ERB) (Patterson *et al.* 1992) is used in Patterson’s model to determine the bandwidth of each cochlear filter. Each filter has an equal amount of overlap with its neighbor filter due to the spacing of the cochlear channels. For the ERB at each center frequency (f_c) of the cochlea channel, Glasberg and Moore (Glasberg and Moore 1990) recommended the following equation.

$$ERB = \frac{f_c}{ErQ} + BW_{min} \quad (3.1)$$

Where BW_{min} is the minimum bandwidth for low-frequency bands, and ErQ is the asymptotic filter quality at high frequencies. These parameters are selected based on with Glasberg and Moore recommends ($ErQ=9.26449$, $BW_{min}=24.7$). In Paterson’s model (Patterson *et al.* 1992), each filter is one ERB band, and the highest and lowest frequencies, as well as the required number of channels, are used to specify the center frequency spacing between channels.

$$f_c = e^{\frac{1}{N}((1:N)(-\log(F_h+ErQ*BW_{min})+\log(F_l+ErQ*BW_{min})))} * (F_h + ErQ * BW_{min}) - (ErQ * BW_{min}) \quad (3.2)$$

Where N represents the number of channels (or) bands, and F_h, F_l represents the high and low frequency of the filter bank respectively. This study considers 16 channels and each channel center frequency is shown in Table 3.1. The filter response of 16 bands filter is shown in Fig. 3.2. The input speech signal first goes through a bank of auditory filters. The 16-channel gammatone filter bank band frequencies are separated logarithmically between 80 and 7562 Hz as shown in Fig. 3.2.

Table 3.1: Center frequencies of the corresponding channel number

Band number(N)	Centre frequency(f_{Nc})	Band number(N)	Centre frequency(f_{Nc})
1	80 Hz	9	1322.3 Hz
2	149 Hz	10	1669.1 Hz
3	233.5 Hz	11	2093 Hz
4	336.8 Hz	12	2612.4 Hz
5	463.3 Hz	13	3247.5Hz
6	618 Hz	14	4024.6 Hz
7	807.3 Hz	15	4975.3Hz
8	1038.9 Hz	16	6138.6 Hz

The response of each filter band was transferred to the Hilbert transform, which gives the imaginary vector values to the given real vector. The analytic signal ($X_a(t)$) was derived by combining the real ($x_r(t)$) and imaginary vectors ($x_i(t)$) of the bands.

$$X_a(t) = x_r(t) + ix_i(t) \quad (3.3)$$

The speech envelope (ENV) was computed from the analytic signal ($X_a(t)$) by using an envelope detector, which consists of a full wave rectifier followed by a low-pass filter

with a cut-off frequency of 50 Hz.

$$ENV(t) = \sqrt{x_r^2(t) + x_i^2(t)} \quad (3.4)$$

The extracted envelope from all sub-bands were modulated by the carrier signal. In this study, the ENV is modulated with the carrier in three signal processing conditions considering different carriers as shown in Fig. 3.1. In a typical acoustic cochlear implant like a sinewave vocoder as shown in Fig. 3.1(a), the ENV is modulated by a sinusoidal carrier. The sinusoidal carrier is extracted from each band center frequency (f_{Nc}). For every band (N), the sinusoidal carrier is defined as follows:

$$C_{N1}(t) = \cos(2\pi * f_{Nc}t) \quad (3.5)$$

where t is length of the band. The modulated speech is the product of envelope ($ENV(t)$) cue and carrier signal $C_{N1}(t)$

$$S_1(t) = ENV_N(t) * C_{N1}(t) \quad (3.6)$$

The speech signal $S_1(t)$ was reconstructed back using the filter bank same as the one used in the analysis stage. Finally, the output of all bands was added separately to get the resultant synthesized speech output of the sinewave vocoder.

The temporal fine structures (TFS) were used as a carrier for modulating envelope in two additional signal processing methods, Full band TFS (Fig. 3.1(b)) and TFS up to 1000 Hz (Fig. 3.1(c)). The TFS information can be derived from the phase of the analytic signal as follows:

$$C_{N2}(t) = \cos(\phi(t)) \quad (3.7)$$

Where $\phi(t)$ is defined as follows:

$$\phi(t) = \tan^{-1} \frac{x_i(t)}{x_r(t)} \quad (3.8)$$

In the full band TFS condition, all the 16 band ENV cues modulated with the corresponding bands TFS

$$S_2(t) = ENV_N(t) * C_{N2}(t) \quad (3.9)$$

The speech signal $S_2(t)$ was reconstructed back using the filter bank same as the one used in the analysis stage. Finally, the output of all bands was added separately to get the resultant synthesized speech output of the full band TFS.

Finally, for TFS up to 1000 Hz condition as shown in Fig. 3.1(c), the lower 8 bands ENV cues were modulated with their corresponding TFS, and the upper 8 bands ENV cues were modulated with their corresponding sinusoidal signal as follows:

$$S_3(t) = \begin{cases} ENV_N(t) * C_{N2}(t) & \text{if } 1 \leq N \leq 8 \\ ENV_N(t) * C_{N1}(t) & \text{if } 8 < N \leq 16 \end{cases} \quad (3.10)$$

The speech signal $S_3(t)$ was reconstructed back using the filter bank same as the one used in the analysis stage. The output of all bands was added separately to get the resultant synthesized speech output of the TFS up to 1000 Hz condition.

3.2.1 Experimental design and results

3.2.1.1 Quick Speech In Noise (QuickSIN)

The Quick speech in noise (Quick-SIN) ([Avinash et al. 2010](#)) test was conducted to evaluate the performance of three signal processing methods. Speech-in-Noise tests are designed to mimic real-life circumstances. In each list, there are 7 sentences, and in each sentence, there are 5 keywords. Thus based on the response of the subject, the score was given between 0 to 5. In this study, each sentence was presented to the listeners at different SNR levels. The variance of the three methods was measured using the speech recognition threshold in noise (SRTN). The SRTN was calculated using Finney's (1952) Spearman Karber Equation given by:

$$SRTN = i + (d/2) - (d * \frac{C}{W}) \quad (3.11)$$

Where i = initial presenting SNR

d = step size (+5dB)

W = identified keywords per decrement with SNR

C = number of key words correctly identified

3.2.1.2 Subjects

Five normal hearing (NH) persons (average age: 29.5 years; age ranges from 24 to 32 years) participated in the current study with their self-report, and screening audiometry for hearing of the listeners within 15 dBHL at octave frequencies range 250-8000 HZ.

3.2.1.3 Stimuli

The speech intelligibility test was performed to verify these three methods through QuickSIN standard Kannada sentences (Avinash *et al.* 2010). The QuickSIN test comprises 7 different lists, with each list having 7 sentences, and each sentence having 5 keywords. To the target speech, a four-talker babble noise was added and the processed speech was given at 7 different SNR levels (+20 dB, +15 dB, +10 dB, +05 dB, 0, -05 dB, -15 dB). The signals pre-processed in MATLAB were given to NH volunteers via Sennheiser HD280pro headphones. A practice trial has been given to all participants to avoid potential learning effects. Once the individuals became familiar with the task, they were subjected to the actual perceptual test. The sentence list for each signal processing condition was randomized for each participant. Participants were instructed to listen carefully to the target speech and repeat them in written form. If the participants were unsure about the sentences, they were permitted to guess. Their responses were evaluated based on the number of correctly identified words in each sentence. A score of 1 was given to each correctly identified keyword. Finally, the total score was counted for three signal processing conditions in each SNRs. The mean scores were converted to proportion correct (PC) scores (normalized between 0 to 1) for fitting the psychometric function. The proportion correct score represents the speech intelligibility. The psychometric function shows speech intelligibility (proportion correct) concerning the processed speech signal SNR levels, which can be defined as

$$f(x) = \frac{1}{1 + e^{\frac{-(x-SRTN)}{m}}} \quad (3.12)$$

Where x is the processed speech signal SNR level and m is the slope of the plot.

The proportion correct score for three signal processing methods was plotted on the Gaussian psychometric function as shown in Fig. 3.3. The speech recognition threshold in noise (SRTN) was measured with respect to the midpoint of the proportion correct, which represents the minimum SNR required for 50% speech intelligibility

(or) proportion correct.

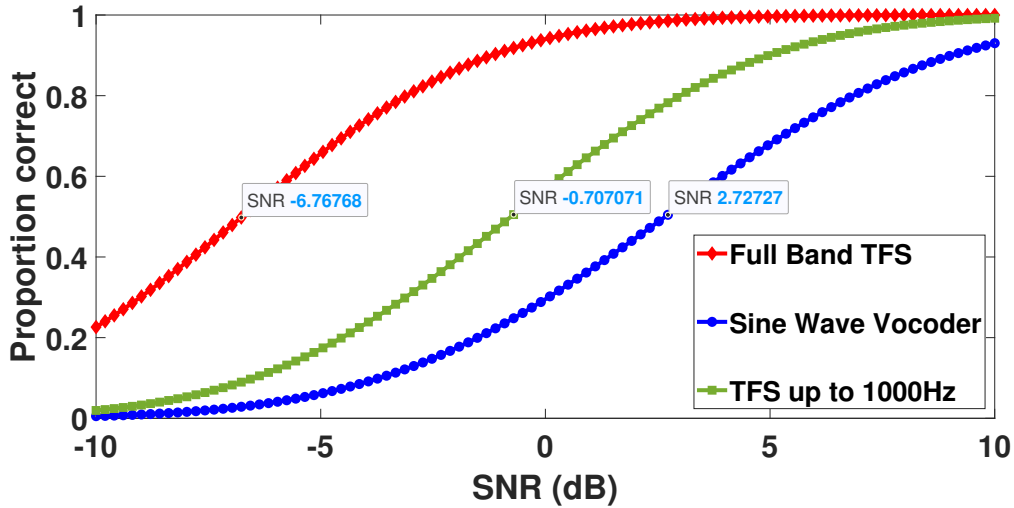


Figure 3.3: SV Vs Full band TFS

The blue curve in Fig. 3.3, represents the Sinewave vocoder, which requires 2.27 dB SNR for 50% speech intelligibility. The Full band TFS represents the red curve, which requires less SNR (-6.767 dB) for 50% speech perception because all band envelopes were modulated with their original signal TFS. Therefore, the full-band TFS condition speech quality is similar to the speech perceived by normal-hearing people. Finally, the required SNR for TFS up to 1000 Hz was -0.707 dB, which represents the green curve in Fig. 3.3. Hence, this result depicts that the sine-wave vocoder requires high SNR for a minimum of 50% speech perception in noise when compared to TFS conditions. This shows that the TFS plays a significant role in CIs for better speech perception in noise.

This study evaluated the speech perception in noise for three models: sinewave vocoder, full band TFS and TFS up to 1000 Hz. In the sinewave vocoder, the envelope is modulated with a sinusoidal signal, and in the full-band TFS model, the envelope is modulated with TFS. In the TFS up to 1000 Hz method, temporal envelopes of the lower eight bands were modulated with TFS, and the upper 8 band envelopes were modulated with a sinusoidal signal. The speech recognition score of the TFS up to 1000 Hz strategy gives a better response than the sinewave vocoder whereas gives a significantly low performance than full band TFS in the case of speech recognition with noise. The full band TFS method was used here to observe the significance of the fine structure for speech recognition in noise, but it is not suitable in real-time

CIs, due to the frequency listening threshold of CI users. TFS coding up to 1000 Hz gives a better speech perception in noise within the frequency listening threshold of some CI users. However, in most CI recipients the temporal encoding saturates by 300 Hz.

From the literature survey, the coding limit of most CI users is 300 Hz, and some well-performing candidates can allow up to 1000 Hz. Hence, the optimum TFS coding should be below 1000 Hz, but still, there are some limitations in 1000 Hz TFS coding, which may cause limited harmonics availability within 1000 Hz. Due to this, two additional cut-off frequencies, TFS up to 2000 Hz and TFS up to 600 Hz were taken into consideration in the next section to examine the effect of harmonicity on speech intelligibility with TFS cut-off frequency.

3.3 Encoding frequency compressed TFS within the neuro-physiological limitations (within 300Hz) of the CI users

In this section, the perceptual benefits of various TFS cut-off frequencies in CI simulation are discussed. Based on the findings, an algorithm is proposed to encode the compressed TFS within the neuro-physiological limitations of the CI users. The perceptual benefits of TFS cut-off frequencies and compressed TFS conditions are compared with the sinewave vocoder.

3.3.1 TFS cut-off frequencies

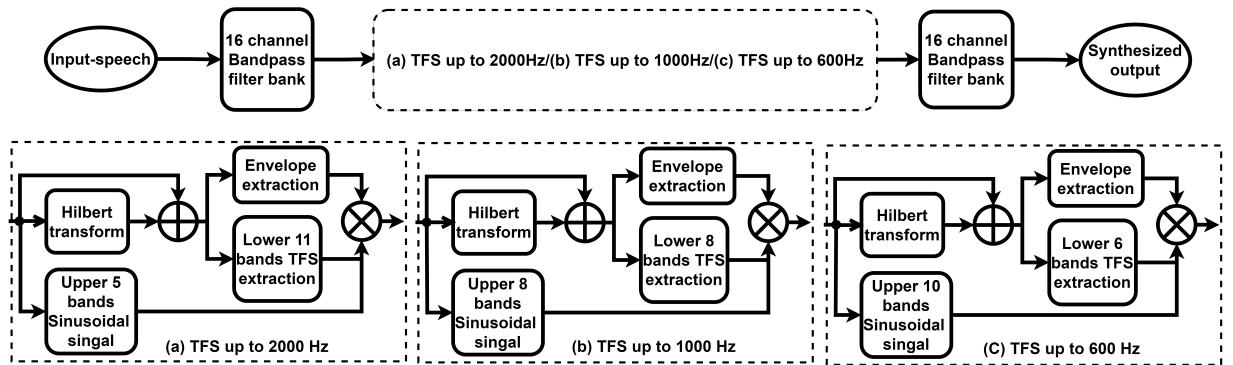


Figure 3.4: Block diagram representation of TFS cut off frequencies

The TFS was extracted from a few low-frequency bands for different TFS cut-off frequency conditions as shown in Fig. 3.4. To demonstrate the impact of TFS cut-off frequency on speech intelligibility in CI simulation, this study has chosen three TFS cut-off frequencies: TFS 2000 Hz, TFS 1000 Hz, and TFS 600 Hz. For the 2000Hz TFS cut-off frequency (Fig. 3.4(a)), TFS up to 11 bands (2093) were extracted. For the remaining 12 to 16 bands, the sinusoidal signal was extracted from the center frequency of the corresponding bands. Similarly, for TFS 1000 Hz condition, 8 bands (1038 Hz) of TFS and the remaining 8 bands of the sinusoidal signal were extracted. Finally, for TFS 600 Hz condition the fine structure was extracted for the lower 5 bands (618 Hz) as shown in Fig. 3.4(c), and for the remaining bands, the sinusoidal signals were used for modulating the envelope. The Sinusoidal signal derived from the center frequency (f_c) of the corresponding band is ($C_{N1}(t)$) derived in equation 3.5.

The input speech signal ($x(t)$) first goes through a bank of auditory filters. A common model of cochlear filtering is the gammatone filter bank (De Boer and De Jongh 1978). A 16-channel gammatone filter bank (Patterson *et al.* 1987) with centre frequencies separated logarithmically between 80 and 7562 Hz are used. The response of gamma tone filter is mathematically expressed as

$$G(t) = \begin{cases} t^{n-1} e^{-2\pi ERBt} * \cos(2\pi f_c t) & ; t \geq 0 \end{cases} \quad (3.13)$$

Where n denotes the filter order and ERB denotes the equivalent rectangular bandwidth that grows as the centre frequency f_c grows. The 16 channel gamma-tone filter bank response is shown in Fig. 3.2. The output of the gammatone filter is

$$S_N(t) = x(t) * G(t) \quad (3.14)$$

Where N stands for channel (or) band number. This study uses the analytic signal to extract the envelope and TFS in each band. The Hilbert transform of the filtered output $S_N(t)$ is $S_a(t)$

$$S_a(t) = S_r(t) + iS_i(t) \quad (3.15)$$

The envelope of the analytic signal ($S_a(t)$) is computed using a rectifier following a low pass filter with a cut-off frequency of 50 Hz, where the magnitude of the analytic

signal is the output of the rectifier

$$ENV_N(t) = \sqrt{S_r^2(t) + S_i^2(t)} \quad (3.16)$$

The TFS information derived from the phase of the analytic signal is defined as

$$C_{N2}(t) = \cos(\phi(t)) \quad (3.17)$$

Where

$$\phi(t) = \tan^{-1} \frac{S_i(t)}{S_r(t)} \quad (3.18)$$

For all TFS cut-off frequency conditions, 16 ENVs were extracted but the modulating carrier is different for each condition. The modulated output of three conditions is defined as follows:

The modulated output of TFS up to 2000 Hz is

$$S_{2000TFS}(t) = \begin{cases} ENV_N(t) * C_{N2}(t) & \text{if } 1 \leq N \leq 11 \\ ENV_N(t) * C_{N1}(t) & \text{if } 11 < N \leq 16 \end{cases} \quad (3.19)$$

The modulated output of TFS up to 1000 Hz is

$$S_{1000TFS}(t) = \begin{cases} ENV_N(t) * C_{N2}(t) & \text{if } 1 \leq N \leq 8 \\ ENV_N(t) * C_{N1}(t) & \text{if } 8 < N \leq 16 \end{cases} \quad (3.20)$$

The modulated output of TFS up to 600 Hz is

$$S_{600TFS}(t) = \begin{cases} ENV_N(t) * C_{N2}(t) & \text{if } 1 \leq N \leq 5 \\ ENV_N(t) * C_{N1}(t) & \text{if } 5 < N \leq 16 \end{cases} \quad (3.21)$$

The synthesized speech was reconstructed using the same filter bank as the one used in the analysis stage. Finally, the output of all bands was added to get the resultant synthesized speech output of the three conditions.

3.3.2 Pitch synchronous overlap add algorithm (PSOLA) for TFS shifting

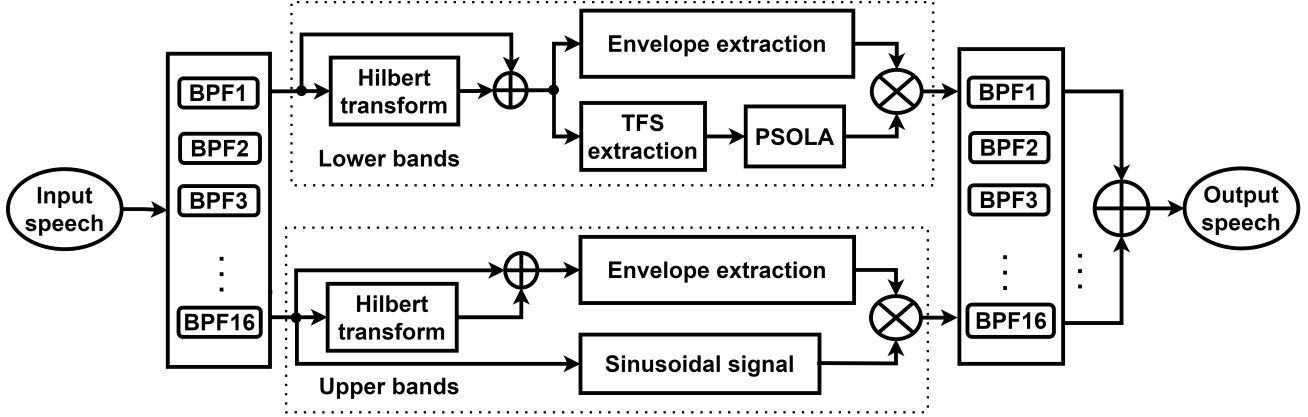


Figure 3.5: Block diagram representation of the proposed speech encoder with PSOLA

The upper limit for CI users to benefit from temporal pitch cues is usually around 300 Hz, which is low when compared to normal hearing people (Zeng 2002). As a result, the TFS will be accommodated within the physiological saturation limits of the cochlear implant while preserving the cues required for speech recognition in noise using the proportional pitch shifting by using pitch synchronous overlap add algorithm (PSOLA) (Moulines and Charpentier 1990, Schnell *et al.* 2000). Pitch shifting alters an audio signal’s pitch up or down without missing any frequency and harmonic information. The spectral envelope (formant positions) is preserved when pitch shifting is performed, which is one of the PSOLA method’s key benefits. Therefore, the extracted TFS ($C_{N_2}(t)$) is transferred to the PSOLA algorithm as shown in Fig. 3.5.

PSOLA is a method based on the decomposition of a signal into a sequence of simple waveforms, where each waveform represents one of the signal’s subsequent pitch periods and their total (overlap-add) reassembles the signal. In PSOLA analysis, the TFS signal is decomposed into a series of elementary waveforms $TFS_i(t)$. This decomposition is achieved through the use of analysis windows $H(t)$ with pitch markers m_i .

$$TFS_i(t) = H(t - m_i) * TFS(t) \quad (3.22)$$

Where m_i is called pitch markers (Peeters 1998). In this study, the analysis window ($H(t)$) is Hanning window.

The analysis pitch-marks (m_i), also known as the time-instants P_i , are established on the voiced segments of speech at a pitch-synchronous rate and on the unvoiced portions at a constant rate. The PSOLA method comprises three key steps:

1. Pitch identification:

The first step is to separate the given TFS signal into small speech segments as shown in Fig. 3.6(c). Then, each frame's autocorrelation sequence is determined. By determining the peaks (PKS) of the autocorrelation sequence and the locations of the peaks (LOC), the pitch period can be calculated.

2. Pitch marking:

On the TFS segment, pitch was marked by marking peak amplitudes, corresponding to peak locations. PSOLA markers m_i must be pitch synchronously placed and close to local maxima. The representation of pitch markers (m_i) in dotted lines is shown in Fig. 3.6 (d).

3. Pitch shifting:

After successful completion of pitch marking of TFS, pitch-shifting by a factor of 0.5 was chosen (range from 0.4 to 0.8) based on the subjective perception and quality of speech within the temporal limitation of TFS coding.

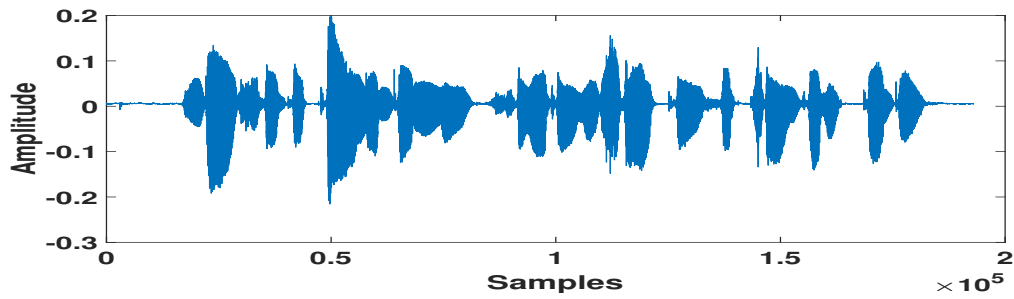
The detailed explanation of the PSOLA processing steps is given in Algorithm 1. The extracted envelope from all sub-bands is modulated by the carrier signal. The TFS was extracted in three different conditions based on TFS cut-off frequency such as TFS up to 2000 Hz, 1000 Hz, and 600 Hz. The modulated speech of three pitch-shifted TFS cut-off frequencies was defined as follows:

The modulated output of pitch shifted TFS up to 2000 Hz is

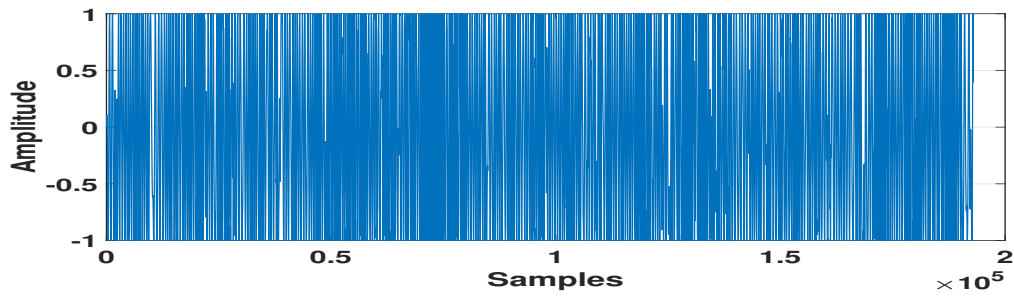
$$S_{P2000TFS}(t) = \begin{cases} ENV_N(t) * TFS_{N_i}(t) & \text{if } 1 \leq N \leq 11 \\ ENV_N(t) * C_{N1}(t) & \text{if } 11 < N \leq 16 \end{cases} \quad (3.23)$$

The modulated output of pitch shifted TFS up to 1000 Hz is

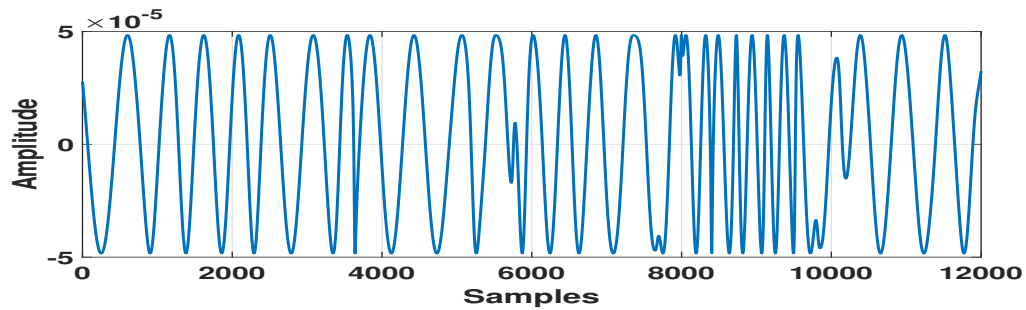
$$S_{P1000TFS}(t) = \begin{cases} ENV_N(t) * TFS_{N_i}(t) & \text{if } 1 \leq N \leq 8 \\ ENV_N(t) * C_{N1}(t) & \text{if } 8 < N \leq 16 \end{cases} \quad (3.24)$$



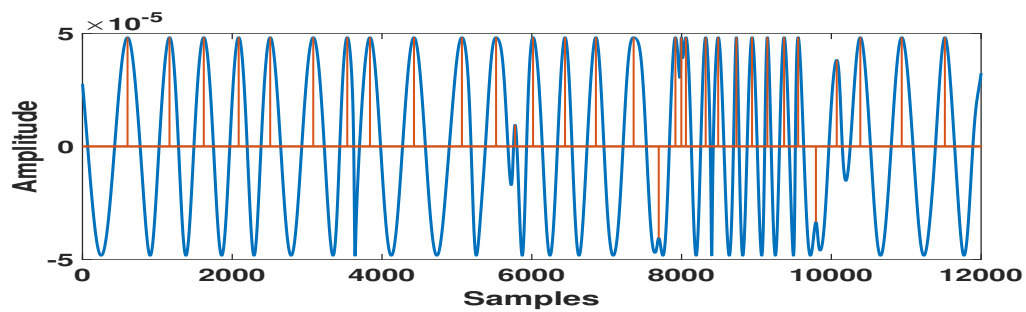
(a) Input speech



(b) Temporal fine structures



(c) Small TFS segment



(d) Pitch marking at peaks of the segment

Figure 3.6: Pitch marking at local maxima of TFS

Algorithm 1: Steps involved in the PSOLA algorithm

Result: Pitch shifted TFS

Input: TFS

Step 1: Pitch identification

1. Select a segment of input signal (TFS)
2. Find peaks (PKS) and locations of the peaks (LOC)
3. Find the peak incidents

$$Peak1_{inci} = \min\left(\frac{Sampling\ frequency}{diff(LOCS)}\right)$$

4. Find the the pitch period

$$Period(T) = LOCS(peak1_{inci} + 1) - LOCS(peak1_{inci})$$

$$Pitch = \frac{1}{T}$$

Step 2: Pitch marking

1. Before pitch mark, locate the peak amplitude locations (LOC).
2. The mark point should have a peak amplitude to constitute a pitch mark.
3. For the upcoming segment pitch mark, after the current segment, the next segment will be reviewed.

Next segment = current segment + number of samples (i) in segment

Step 3: Pitch shifting

1. Pitch-shifting by a factor of 0.5
2. If the first pitch period is greater than the first pitch mark (m), the first pitch mark to be removed
3. Remove the last pitch mark

$$\text{if } m(\text{length}(m)) + T(\text{length}(T)) > \text{length}(\text{input})$$

Step 4: Overlap and add

1. Compute the analysis segment
segment = input(start:end) x Hanning($2 * T_i + 1$)
2. compute the output pitch mark (m_o)

$$m_o = m_o + \frac{T_i}{\beta}$$

3. Overlap and add a segment

$$\text{output}(\text{start}:\text{end}) = \text{output}(\text{start}:\text{end}) + \text{previous segment}$$

$$\text{where } \text{start} = m_o - T_i$$

$$\text{end} = m_o + T_i$$

The modulated output of pitch shifted TFS up to 600 Hz is

$$S_{P600TFS}(t) = \begin{cases} ENV_N(t) * TFS_{N_i}(t) & \text{if } 1 \leq N \leq 5 \\ ENV_N(t) * C_{N1}(t) & \text{if } 5 < N \leq 16 \end{cases} \quad (3.25)$$

The ENV of sub-band signals was then scaled to the respective sub-band signal’s root-mean-square (RMS) levels extracted from the initial band-pass filtering. In each of the 16 bands, the Full band TFS condition was created by multiplying the ENV with the original TFS. In Sine wave vocoder (NO-TFS), the envelope was multiplied with a sinusoidal signal for all the 16 bands. In TFS 2000 Hz with PSOLA condition, lower 11 bands ENVs were modulated by pitch-shifted TFS cues, and remaining 5 bands ENVs were modulated by a sinusoidal signal. Similarly, in TFS up to 1000 Hz with PSOLA condition, lower 8 bands ENVs were modulated with lower 8 channel pitch-shifted TFS cues and upper 8 bands ENVs were modulated by a sinusoidal signal. Finally, in TFS up to 600 Hz with PSOLA condition, lower 6 bands, the ENVs were modulated with pitch-shifted TFS cues and rest of the bands ENVs were modulated by a sinusoidal signal. The speech signal $S_o(t)$ was reconstructed back using the Gamma-tone filter bank same as the one used in the analysis stage. Finally, for each condition, the output of all bands were added separately to the resultant synthesized speech output.

3.4 Experimental Results

3.4.1 Participants

Five native Kannada speakers (2 male and 3 female) aged between 30–40 yrs participated in the current study. Sample considered in the current study fulfills the criteria for the minimum required sample size for the psycho-physical research ([Anderson and Vingrys 2001](#)). Participant’s hearing thresholds were within normal limits (≤ 15 dBHL) at audio-metric octave frequencies ranging from 250 to 8000 Hz. In a quiet listening situation, all of the participants achieved an unprocessed speech identification score of $\geq 80\%$. Prior to their participation in this study, the subjects gave written informed consent in accordance with the Declaration of Helsinki. The study was given approval by the local Ethics Committee (Approval Number:NITK-EC-PhD-392-2021).

3.4.2 Speech Recognition In Noise (SRIN) Test

Table 3.2: Sample procedure of lists presented to NH Volunteers

SNR (dB)	Condition	LISTS
10	Sinewave Vocoder	list1
	Full band TFS	list2
	TFS 2000	list3
	TFS 1000	list4
	TFS 600	list5
	TFS 2000PSOLA	list6
	TFS 1000PSOLA	list7
	TFS 600PSOLA	list8
0	Sinewave Vocoder	list9
	Full band TFS	list10
	TFS 2000	list11
	TFS 1000	list12
	TFS 600	list13
	TFS 2000PSOLA	list14
	TFS 1000PSOLA	list15
	TFS 600PSOLA	list16
-10	Sinewave Vocoder	list17
	Full band TFS	list18
	TFS 2000	list19
	TFS 1000	list20
	TFS 600	list21
	TFS 2000PSOLA	list22
	TFS 1000PSOLA	list23
	TFS 600PSOLA	list24

Perceptual contribution of TFS and the performance of the proposed algorithm are verified through SRIN test using standard pre-recorded Kannada sentences ([Geetha et al. 2014](#)). For each signal processing condition, listener’s SRIN ability was assessed at +10, 0 and -10 dB SNR. Therefore 24 lists (8 conditions x 3 SNRs) of sentences were used for the perceptual experiment. A 4-talker babble was added to the input speech at the desired SNR subjected to signal processing method. The processed speech signals were presented over the Sennheiser HD280pro at most comfortable level. The participants were instructed to focus on the target speech while ignoring the noise in the background. The responses were collected in the written form. All the participants

were familiarized with task prior to the actual experiment by providing the practice trials.

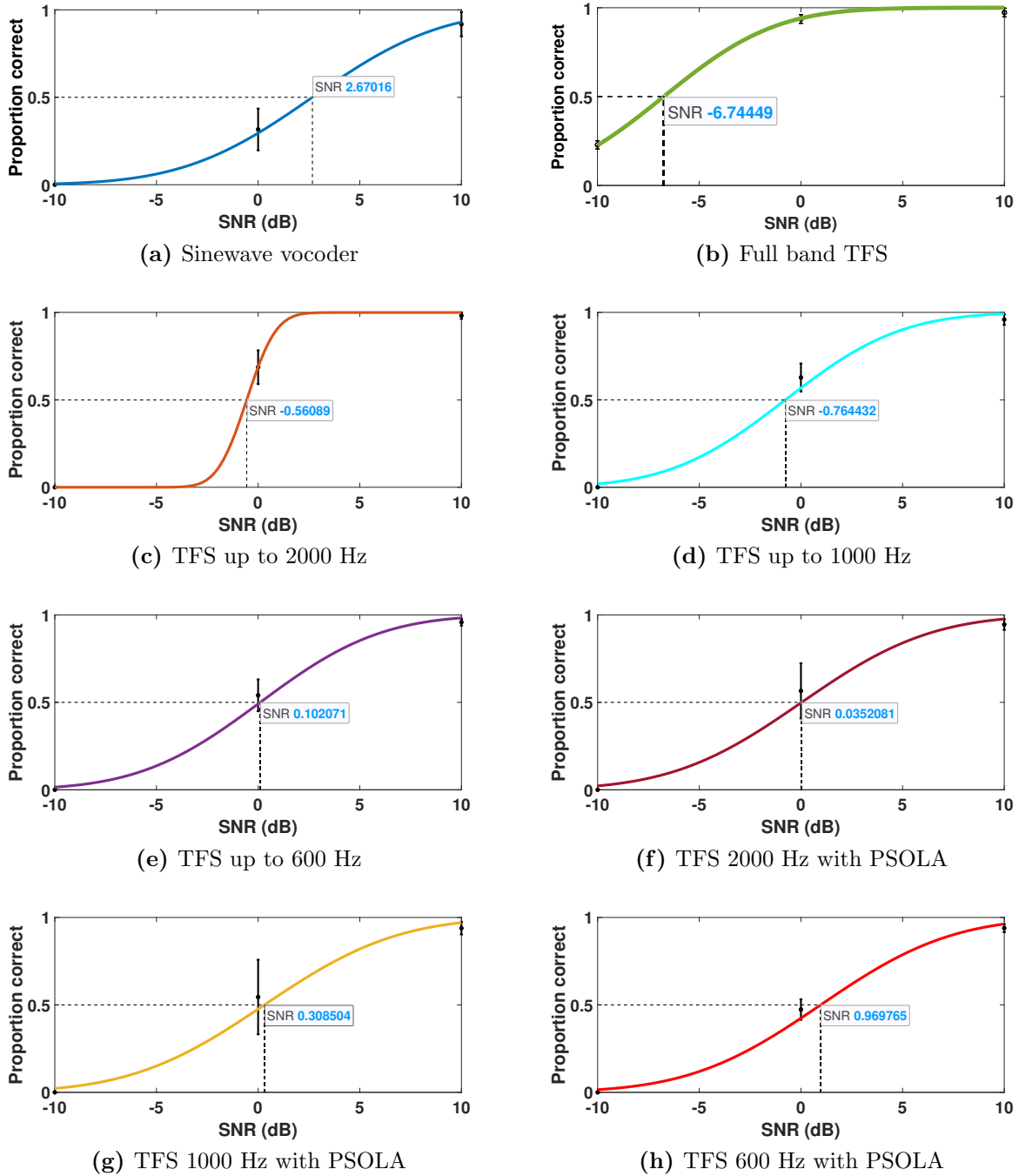


Figure 3.7: Psychometric plots of eight signal processing techniques

Eight conditions were tested with 24 lists, each condition was presented with three lists at 3 different SNR levels (+10 dB, 0 dB, -10 dB) as shown in Table 3.2. In the first session, the Sine wave vocoder and Full band TFS were presented with three SNR levels (+10 dB, 0 dB, -10 dB) for six different lists. In the second session, the cut-off frequencies of TFS 2000 Hz, 1000 Hz, and 600 Hz were given with three SNR levels (+10 dB, 0 dB, -10 dB) for 9 lists. In the third session, the pitch-shifted TFS of 2000 Hz, 1000 Hz, and 600 Hz were given with three SNR levels (+10 dB, 0 dB, -10 dB) for the remaining 9 lists. For different SNRs, the speech recognition score was estimated as the total number of correctly identified keywords. For every correctly identified keyword, a score of one was given. Total correct scores were calculated by counting the number of correctly identified keywords, in each SNR and signal processing conditions. The mean scores were converted to proportion correct (PC) scores (normalized between 0 to 1) for fitting the psychometric function. The proportion correct score represents speech intelligibility.

Table 3.3: Mean speech intelligibility scores (PC) of volunteers

Condition	Mean		
	+10 dB	0 dB	-10 dB
Sinewave vocoder	0.91	0.316	0
Full band TFS	0.97	0.936	0.228
TFS up to 2000 Hz	0.98	0.68	0
TFS up to 1000 Hz	0.958	0.627	0
TFS up to 600 Hz	0.958	0.54	0
2000 Hz TFS with PSOLA	0.945	0.5656	0
1000 Hz TFS with PSOLA	0.93	0.544	0
600 Hz TFS with PSOLA	0.93	0.47	0

A statistical cumulative Gaussian distribution function was fit to the PC data across the SNR for eight signal processing conditions as shown in Fig. 3.7, the speech recognition in noise (SRTN) was measured in each plot to determine the minimum SNR required to achieve 50% speech intelligibility. The TFS cut-off frequency conditions showed better SRTN than the Sinewave vocoder and also TFS with pitch-shifted (including TFS 600 Hz with PSOLA (0.969 dB)) provided better SRTN than Sinewave vocoder (2.67 dB) can be observed in Fig. 3.7(a) & (h). The mean speech recognition scores (proportion correct scores) of the NH volunteers with three SNR levels were measured for different signal processing conditions as shown in the Table 3.3.

3.4.3 Effect of TFS cut off frequency

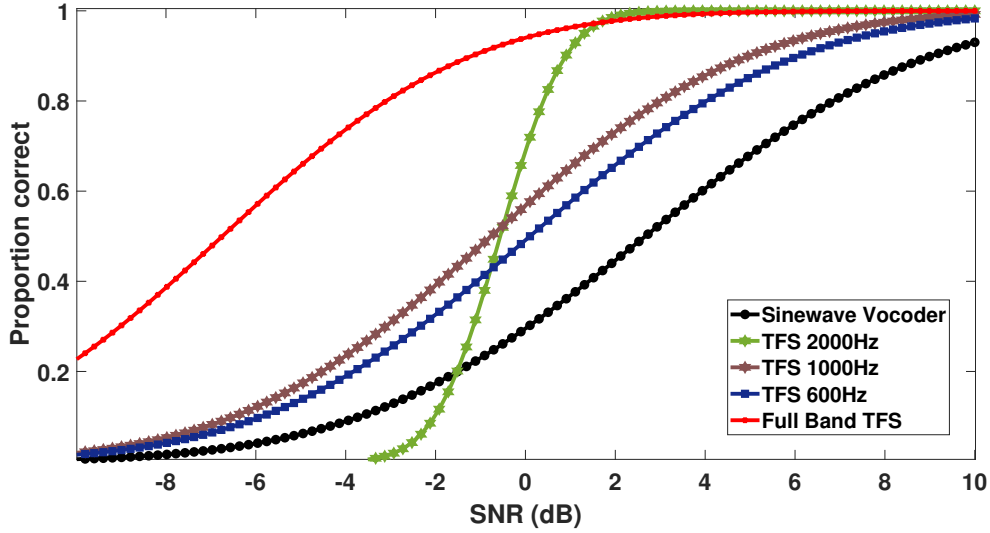


Figure 3.8: Speech intelligibility of sinewave vocoder and TFS cut-off frequencies

The main effect of TFS cut-off frequency on speech recognition score was investigated using Bayesian statistics. For this analysis, the proportion correct data is considered only at +0 dB SNR as the performance at +10 dB SNR has reached the ceiling, and performance at -10 dB SNR is near to the floor for many conditions as shown in the Table 3.3. Speech recognition scores in noise at different signal processing conditions were compared using Bayesian paired sample T tests.

Table 3.4: Bayesian paired sample T-Test between Sinewave vocoder and TFS cut-off frequencies

Measure 1	vs.	Measure 2	BF_{10}
Sinewave vocoder	vs.	TFS up to 2000 Hz	71.66
Sinewave vocoder	vs.	TFS up to 1000 Hz	53.067
Sinewave vocoder	vs.	TFS up to 600 Hz	6.074

The Bayesian paired-sample T-tests were conducted between the sinewave vocoder and three TFS cut-off frequency conditions using the JASP tool. The Paired-Samples T-test compares the mean scores of two conditions for a single group. The procedure computes the differences between the mean scores of the two conditions for each SNR.

and tests whether the average differs from 0. In this test, there are two mutually exclusive hypotheses, such as the null hypothesis (H_0) and the alternative hypothesis (H_1). The null hypothesis (H_0) always favours the condition of equality, whereas the alternative hypothesis (H_1) favours the difference (greater, lesser, or unequal) in conditions. In this study, the Bayesian factor (BF_{10}) indicates favouring H_1 over H_0 , and a higher BF_{10} value indicates a significant difference in TFS cut-off frequency conditions over the sinewave vocoder as shown in Table 3.4 and Fig. 3.8. The PC response for TFS up to 600 Hz is significantly better ($BF_{10}=15.53$) than PC response for without TFS (sine wave vocoder) as shown in Table 3.4.

The statistical analysis revealed that by reducing the Full band TFS to TFS 2000 Hz cut-off frequency, the PC responses significantly reduced ($BF_{10}=13.843$). It has been observed that PC responses have reduced significantly by reducing the Full Band TFS to TFS 1000 Hz cutoff frequency ($BF_{10}=75.146$) and Full band TFS to TFS 600 Hz cut-off frequency ($BF_{10}=55.054$). Similarly, the PC response is reduced by reducing TFS 2000 Hz to cut-off frequency 1000 Hz ($BF_{10}=1.035$), and TFS 2000 Hz to cut-off frequency 600 Hz, the PC response is reduced ($BF_{10}=9.807$). Finally, the PC response is reduced by reducing TFS 1000 Hz to 600 Hz ($BF_{10}=3.539$).

3.4.4 Effect of TFS pitch shifting

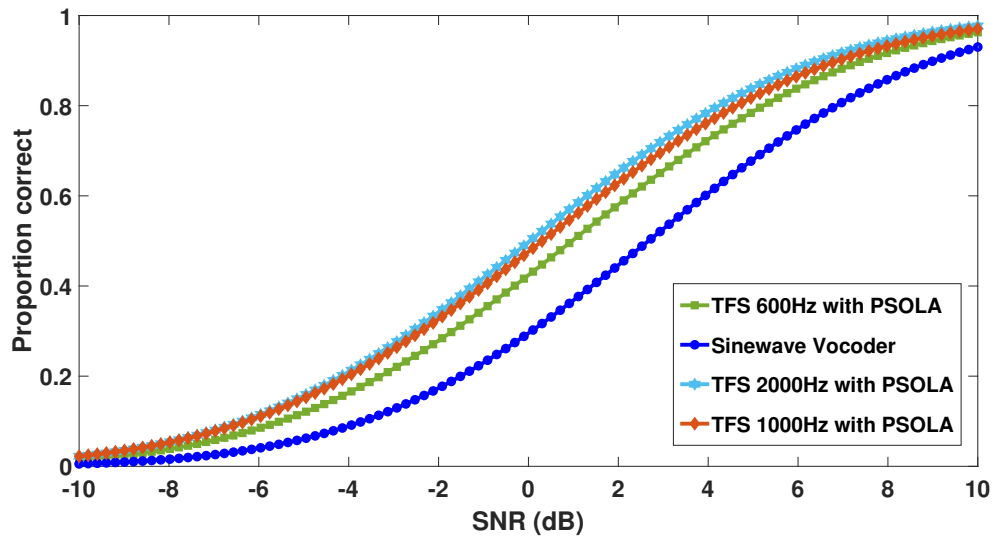


Figure 3.9: Speech intelligibility of sinewave vocoder and pitch shifted TFS

Table 3.5: Bayesian paired sample T-test between sinewave vocoder and pitch shifted TFS

Measure 1	vs.	Measure 2	BF_{10}
Sinewave vocoder	vs.	pitch-shifted TFS up to 2000 Hz	15.537
Sinewave vocoder	vs.	pitch-shifted TFS up to 1000 Hz	2.181
Sinewave vocoder	vs.	pitch-shifted TFS up to 600 Hz	1.54

The TFS pitch information was shifted from 2000 Hz, 1000 Hz, and 600 Hz to 1000 Hz, 500 Hz, and 300 Hz respectively with pitch shifting factor ($\beta=0.5$) using PSOLA. The PC scores at these cut-off frequencies with shifting were compared against the Sinewave vocoder as shown in Fig. 3.9. Fig. 3.9 shows that three pitch-shifted TFS conditions show significant improvement in terms of speech intelligibility (PC) compared to sinewave vocoder.

The Bayesian paired-sample T-tests were conducted between the sinewave vocoder and three pitch shifted TFS conditions as shown in Table 3.5. The statistical analysis revealed that the TFS 2000 Hz pitch-shifting by a factor of 0.5 has significantly improved the performance compared with the Sine wave vocoder ($BF_{10}=71.664$). Similarly, the Bayesian paired sample T-test revealed that the PC response for TFS up to 1000 Hz has been improved due to pitch-shifting by a factor of 0.5 compared with Sine wave vocoder ($BF_{10}=2.181$). Finally, TFS up to 600 Hz with pitch-shifting by a factor of 0.5 has a noticeable improvement in PC response compared with Sine wave vocoder ($BF_{10}=1.54$).

3.5 Discussion

Results of the current study is in consonance with previous studies reporting improvement in speech recognition in noise with addition of TFS (Apoux *et al.* 2015, Hazrati *et al.* 2015). The TFS carries the essential cue for stream segregation such as F0 and harmonics more efficiently than the temporal envelope. Hence, the addition of TFS improves speech recognition in noise by segregating target speech and interfering noise into two separate streams. Apoux *et al.*, suggested that addition of TFS would help in stream segregation by indicating the auditory system that, the temporal envelopes are being carried by different carriers.

Decreasing the TFS cut-off frequency from full band TFS to 2000 Hz, 1000 Hz

and 600 Hz, significantly reduced the speech recognition ability in noise. Nevertheless, the scores were better compared to sinewave vocoded speech. Improvement with speech recognition in noise even while restricting the TFS to low frequency bands can be attributed to super-additive mechanisms (Micheyl and Oxenham 2012). Super additive models assume a form of synergistic interaction between the low passed and high passed information. The segregation cues available in the TFS of the low frequency bands might synergistically interact with ENVs in high frequency to improve the perception. On the other hand, speech recognition scores in noise, decreased with decrease in TFS cut-off frequency. With a lower cut-off frequency, there are fewer harmonics available, which could be the reason for the reduction in speech recognition scores. Harmonics play more important role than F_0 in the perception of pitch and pitch mediated stream segregation. The reduction in the TFS cut-off frequency would reduce the number of harmonics in the encoded speech. This results in reduced pitch perception strength and leading to weak stream segregation.

An important finding of the current work is that the speech recognition in noise is improved with the addition of frequency compressed TFS. In this work, proportional frequency compression was implemented through PSOLA. Pitch shift due to the frequency compression is known to affect speech intelligibility. However, in the current study pitch-shifted speech was more intelligible than sine wave vocoder. Also for a similar cut-off frequency, speech recognition scores with and without frequency compression were not greatly different. These findings favor the use of frequency compression as means to encode TFS within the temporal coding limit. The proportional frequency compression can preserve the harmonicity, and deliver more harmonics within the temporal coding limits of CI. Earlier SSE, HSSE, and CSSS attempted to encode harmonics via TFS. Nevertheless, the SSE failed in preserving the harmonicity, and the benefit of HSSE and CSSS is limited by the available harmonics within the temporal coding limit. Hence, compressing TFS through PSOLA might overcome the limitation of the above strategies. Thus, the proportional frequency compressed TFS has potential implications in the speech coding strategies of the cochlear implant.

3.6 Summary

This chapter, investigated the application of proportional frequency compression method to encode the TFS within the temporal encoding limits of cochlear implants. The mean

speech recognition scores with frequency compressed 600 Hz TFS were better than sine wave vocoder, indicating that frequency compressed TFS seems to have a positive outcome. Hence this study recommends that future CI signal processing strategies can consider implementing the pitch shifting in the signal processing algorithm. However, the results of the current study should be interpreted with the caution as the target speech was spoken by a female speaker. Future studies should tap the male speaker as well. Also, neuro-physiological studies focusing on pitch coding might be considered in the future to physiologically validate the concept. One of the major limitations of the current study is that the pitch perceived by the cochlear implant listener may not be actual pitch. However, this limitation can be ignored as long as the speech recognition improves.

Chapter 4

DESIGN OF AN EFFICIENT NOISE REDUCTION METHOD TO IMPROVE SPEECH RECOGNITION IN CIs

A novel pre-processing method to improve speech Intelligibility in noise is proposed and tested using the acoustic simulations of cochlear implants. The proposed noise reduction technique aims to minimize the mean square error (MSE) between the temporal envelopes of the enhanced speech and clean speech, making suitable for CI applications. This study provides an analysis of the theoretical derivation of the noise suppression function and also the performance evaluation using objective and subjective tests. The effectiveness of the proposed method was objectively evaluated using the speech-to-reverberation modulation energy ratio (SRMR-CI) and extended short time objective intelligibility (ESTOI). Additionally, speech recognition through the acoustic simulations of the cochlear implant was done for the subjective evaluation. Performance of the proposed method was compared with the Wiener filter (WF) and sigmoidal functions. The sinewave vocoder was used to simulate the cochlear implant perception.

4.1 Introduction

The speech recognition capability of the CI users in quiet situation is satisfactory. However, their performance in noisy environments is suboptimal (Nie *et al.* 2004)(Chiea *et al.* 2021)(Remus and Collins 2005). CIs recipients are less likely to identify speech in the existence of background noise than people with normal hearing (NH). CI user’s speech recognition scores are reduced from 60 to 30% when signal-to-noise ratios are low (Spahr *et al.* 2007). However, individuals with cochlear implants require a 25dB higher SNR to recognize a minimum 50% of the target speech given in the background talker noise (Turner *et al.* 2004)(Hast *et al.* 2015). These findings indicate that noise reduction strategies in CIs are a critical link in the signal processing pipeline because they help users to maintain good speech intelligibility even in noisy conditions.

A variety of noise reduction (NR) strategies have been proposed to enhance speech intelligibility (SI), voice quality, and hearing comfort in poor listening situations to overcome issues with speech perception. The goal of NR is to remove as much noise as feasible from a noisy mixture while keeping the target signal distortions to a minimum. Time-frequency masking (TFM) is a type of NR technique frequently used in hearing aids and CIs (Mauger *et al.* 2012)(Hazrati *et al.* 2013). The ideal binary mask (IBM) (Chen *et al.* 2006) (Koning *et al.* 2014) and the Wiener Filter (WF) (Hazrati *et al.* 2013) are the most common methods (Koning *et al.* 2018). When it comes to CI applications, general-purpose masks have their limitations (Henry *et al.* 2021).

CI users are generally encouraged to use suppression functions that are more aggressive than those used in hearing aids or those with normal hearing (Mauger *et al.* 2012) (Hersbach *et al.* 2012) (Mourao *et al.* 2020). When the SNR is above a specified threshold value, IBM preserves the time-frequency points of the noisy signal and suppresses the remaining time-frequency points. Unlike IBM, WF is the method for providing masks with continuous weights. Van Dijk *et al.* 2012, Madhu *et al.* 2012, Koning *et al.* 2018 state that in auditory prosthetics such as hearing aids and CIs, WF has an improvement over the IBM approach. The WF method provides a path to reduce the mean square error between the target and the estimated signals.

In recent studies, machine learning techniques have been used in CIs to reduce noise (Lai *et al.* 2016) (Wang *et al.* 2020) (Tseng *et al.* 2020). WF-based techniques are commonly used for getting the desired target speech with the supervised learning process. Even though they provide attractive results, their accuracy is strongly corre-

lated with the size of data sets for speech and noise, as this can result in over-fitting of the data (training), which may limit their ability to generalize to various acoustic environments.

This study presents a technique for reducing noise in acoustic simulations of CIs. Most cochlear implants encode temporal envelopes. Hence, the current method intends to minimize the MSE between the estimated and target signal's squared envelopes. Simulations reveal that the proposed method (PM) performs better than the WF mask with the minimum MSE. In the current study, the performance of the PM was tested on the acoustic simulations of CI using sinewave vocoder (Crew and Galvin III 2012)(Mesnildrey *et al.* 2016). In the field of CI research, sinewave vocoders frequently serve to replicate some of the characteristics of CI processing (Loizou 2006). Here, the proposed noise reduction method and the traditional single microphone noise reduction method (i.e., WF) are compared in terms of speech recognition performance. This work aims to examine the noise suppression capacity of NR methods under various challenging circumstances. Test data for evaluation is synthesized using two noises with lower SNR levels. For confirming the effectiveness of PM on normal speech, this study uses an objective measurement (the extended short-time objective intelligibility (ESTOI) (Jensen and Taal 2016). The ESTOI measure has shown to be highly accurate in predicting speech intelligibility under many conditions of degradation (Jensen and Taal 2016). To evaluate the performance further, the study uses vocoder speech to conduct a listening experiment with normal hearing (NH) subjects. Psychoacoustic experiments with NH people indicated that the proposed method yields higher speech intelligibility in a wide variety of SNRs.

The following sections of the chapter are organized as follows: In Section 4.2 noise reduction methodology of the proposed method is explained. In Section 4.4, discuss the simulation results of the perceptual and objective evaluations. Section 4.5 of this work provides a discussion. Summary of the chapter is provided in section 4.6.

4.1.1 Noise reduction methodology

The typical noise reduction method applied to the CI simulator is as shown in Fig. 4.1. The general noise additive method defines the noisy speech signal as $y(n) = x(n)+b(n)$, where $b(n)$ is the additive noise, and $x(n)$ is the target speech signal. The spectral components are calculated by Fast Fourier Transform (FFT) with two times window length, with the window length is 20 msec.

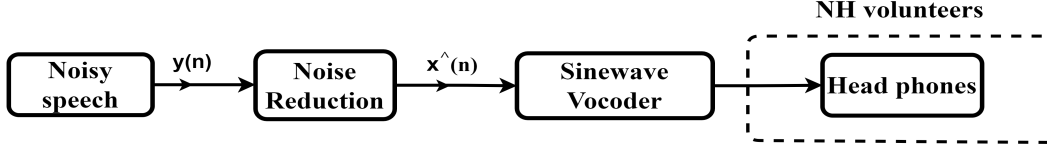


Figure 4.1: Block diagram representing noise reduction and vocoder-based simulation of cochlear implants

The ω^{th} spectral component of noisy short time speech frame (τ_s) can be defined by

$$Y(\tau_s, \omega) = X(\tau_s, \omega) + B(\tau_s, \omega) \quad (4.1)$$

where $X(\tau_s, \omega)$, $B(\tau_s, \omega)$ represents spectral components of clean and noise respectively. The enhanced spectral speech segment $\hat{X}(\tau_s, \omega)$ can be estimated as follows:

$$\hat{X}(\tau_s, \omega) = Y(\tau_s, \omega) \cdot W(\tau_s, \omega) \quad (4.2)$$

where $W(\tau_s, \omega)$ is the noise reduction filter coefficient vector. There are many methods for finding the coefficient vector. They define the filter coefficients based on the functions of noisy speech SNR estimations. The Wiener filter (WF) is an example of a time-frequency mask that has been effectively applied to CI. The WF (Plapous *et al.* 2006) can be defined as

$$W(\tau_s, \omega) = \frac{\gamma(\tau_s, \omega)}{1 + \gamma(\tau_s, \omega)} \quad (4.3)$$

where $\gamma(\tau_s, \omega)$ is an a priori SNR defined as follows:

$$\gamma(\tau_s, \omega) = \frac{E[X^2(\tau_s, \omega)]}{E[B^2(\tau_s, \omega)]} \quad (4.4)$$

in which $E[X^2(\tau_s, \omega)]$, and $E[B^2(\tau_s, \omega)]$ represents clean speech and noise instantaneous powers respectively. $E[]$ represent the expected value operator. From the decision direct approach method (Plapous *et al.* 2004), a priori SNR, is defined as

$$\gamma(\tau_s, \omega) = \alpha \frac{E[X^2(\tau_s - 1, \omega)]}{E[B^2(\tau_s, \omega)]} + (1 - \alpha) \text{Max} \left[\frac{E[Y^2(\tau_s, \omega)]}{E[B^2(\tau_s, \omega)]} - 1, 0 \right] \quad (4.5)$$

Where α is the weighting factor. For better performance of WF, the recommended α value is 0.98.

Similarly, the sigmoid function (Hu *et al.* 2007) employed for noise reduction successfully used for CIs, is defined as follows

$$g(\tau_s, \omega) = e^{-2/\gamma(\tau_s, \omega)} \quad (4.6)$$

(a) Estimation of envelope and phase: The noisy speech signal is windowed, after which the short-time Fourier transform (STFT) is applied. The estimated envelope is the absolute value of its STFT signal, defined as

$$Y_a(\tau_s, \omega) = Y_r(\tau_s, \omega) + iY_i(\tau_s, \omega) \quad (4.7)$$

$$Envelope = \sqrt{Y_r^2(\tau_s, \omega) + Y_i^2(\tau_s, \omega)} \quad (4.8)$$

phase information defined as

$$\phi(\tau_s, \omega) = \tan^{-1} \frac{Y_i(\tau_s, \omega)}{Y_r(\tau_s, \omega)} \quad (4.9)$$

4.2 The proposed noise suppression function

This section introduces a novel optimization framework proposed for obtaining the noise suppression function and calculating the noise power. The proposed suppression function for noise reduction from the minimization of MSE between the desired speech and its enhanced speech, at each spectral band, is given by

$$J(\tau_s, \omega) = \mathbb{E}[e^2(\tau_s, \omega)] \quad (4.10)$$

where $e(\tau_s, \omega)$ is the error between desired speech and its enhanced speech envelope, given by

$$e(\tau_s, \omega) = |X_a(\tau_s, \omega)|^2 - |\hat{X}_a(\tau_s, \omega)|^2 \quad (4.11)$$

The estimated speech $\hat{X}_a(\tau_s, \omega)$ can be expressed as

$$\hat{X}_a(\tau_s, \omega) = V(\tau_s, \omega).Y(\tau_s, \omega) \quad (4.12)$$

where $V(\tau_s, \omega)$ are filter coefficients.

$$e^2(\tau_s, \omega) = (|X(\tau_s, \omega)|^2 - |V(\tau_s, \omega).Y(\tau_s, \omega)|^2)^2 \quad (4.13)$$

By using the above expression (4.13), the equation (4.10) can be written as follows

$$J(\tau_s, \omega) = \mathbb{E}[(|X(\tau_s, \omega)|^2 - |V(\tau_s, \omega).Y(\tau_s, \omega)|^2)^2] \quad (4.14)$$

Assuming that $X(\tau_s, \omega)$ and $B(\tau_s, \omega)$ both have zero mean and are independent of each other (Lu and Loizou 2010), (4.14) can be written as

$$\begin{aligned} J(\tau_s, \omega) = & E [|X_a(\tau_s, \omega)|^4] + \\ & |V(\tau_s, \omega)|^4 E [|X_a(\tau_s, \omega)|^4] \\ & + 4|V(\tau_s, \omega)|^4 E [|X_a(\tau_s, \omega)|^2] E [|B_a(\tau_s, \omega)|^2] \\ & + |V(\tau_s, \omega)|^4 E [|B_a(\tau_s, \omega)|^4] \\ & - 2|V(\tau_s, \omega)|^2 E [|X_a(\tau_s, \omega)|^4] \\ & - 2|V(\tau_s, \omega)|^2 E [|X_a(\tau_s, \omega)|^2] E [|B_a(\tau_s, \omega)|^2] \end{aligned} \quad (4.15)$$

For minimizing, the above equation (4.15) can be differentiated with respect to $V(\tau_s, \omega)$ and equated to zero (Hjorungnes and Gesbert 2007) which gives

$$\begin{aligned} |V(\tau_s, \omega)|^2 (E [|X_a(\tau_s, \omega)|^4] + 4E [|X_a(\tau_s, \omega)|^2] E [|B_a(\tau_s, \omega)|^2] + E [|B_a(\tau_s, \omega)|^4]) \\ = E [|X_a(\tau_s, \omega)|^4] + E [|X_a(\tau_s, \omega)|^2] E [|B_a(\tau_s, \omega)|^2] \end{aligned} \quad (4.16)$$

The above equation can be written as

$$\left| V(\tau_s, \omega) \right| = \sqrt{\frac{E [|X_a^4(\tau_s, \omega)|] + E [|X_a(\tau_s, \omega)|^2] E [|B_a(\tau_s, \omega)|^2]}{E [|X_a^4(\tau_s, \omega)|] + E [|B_a^4(\tau_s, \omega)|] + 4E [|X_a(\tau_s, \omega)|^2] E [|B_a(\tau_s, \omega)|^2]}} \quad (4.17)$$

Let $\sigma_{ax}^2(\tau_s, \omega)$, $\sigma_{ab}^2(\tau_s, \omega)$ represent acoustic clean and noise signal powers respectively, which can be written as

$$\sigma_{ax}^2(\tau_s, \omega) = 2\sigma_x^2(\tau_s, \omega) = E[X_a^2(\tau_s, \omega)] \quad (4.18)$$

$$\sigma_{ab}^2(\tau_s, \omega) = 2\sigma_b^2(\tau_s, \omega) = E[B_a^2(\tau_s, \omega)] \quad (4.19)$$

Equation 4.17 can be rewritten by substituting equations 4.18 and 4.19.

$$\left| V(\tau_s, \omega) \right| = \sqrt{\frac{E[|X_a^4(\tau_s, \omega)|] + \sigma_{ax}^2(\tau_s, \omega) \cdot \sigma_{ab}^2(\tau_s, \omega)}{E[|X_a^4(\tau_s, \omega)|] + E[|B_a^4(\tau_s, \omega)|] + 4\sigma_{ax}^2(\tau_s, \omega) \cdot \sigma_{ab}^2(\tau_s, \omega)}} \quad (4.20)$$

For the analytic speech signal, the fourth order expected value in (4.20) could be expressed as (4.21) in accordance with (Chiea *et al.* 2021)

$$E[|X_a^4(\tau_s, \omega)|] = 2\sigma_x^2(\tau_s, \omega) \cdot 2\sigma_x^2(\tau_s, \omega) \quad (4.21)$$

Similarly, the fourth-order noise power can be written as follows.

$$E[|B_a^4(\tau_s, \omega)|] = 2\sigma_b^2(\tau_s, \omega) \cdot 2\sigma_b^2(\tau_s, \omega) \quad (4.22)$$

Estimation of noise power: The voice activity detection method is used to decide whether the input signal contains noise or speech based on the speech presence probability (SPP), with usual probability threshold (PTH) between 0 to 1. In this work, the SPP greater than or equal to 0.6 for speech presence is considered based on the pilot study (Sohn *et al.* 1999)(Martin 2001). SPP of less than 0.6 is considered as the noise. The noise power spectral density is calculated using the following typical recursive relation (Loizou 2006)

$$E[B_a^2(\tau_s, \omega)] = \lambda E[B_a^2(\tau_s - 1, \omega)] + (1 - \lambda) E[|Y_a(\tau_s, \omega)|^2] \quad (4.23)$$

λ is the smoothing factor whose value ranges from 0 to 1. Equation 4.20 can be rewritten by substituting equations 4.21 and 4.22.

$$\left| V(\tau_s, \omega) \right| = \sqrt{\frac{4(\sigma_x^2(\tau_s, \omega))^2 + 4\sigma_x^2(\tau_s, \omega) \cdot \sigma_b^2(\tau_s, \omega)}{4(\sigma_x^2(\tau_s, \omega))^2 + 4(\sigma_b^2(\tau_s, \omega))^2 + 16\sigma_x^2(\tau_s, \omega) \cdot \sigma_b^2(\tau_s, \omega)}} \quad (4.24)$$

In the above equation, numerator and denominator terms are divided by $4(\sigma_b^2(\tau_s, \omega))^2$; then, the equation can be written as

$$V(\tau_s, \omega) = \sqrt{\frac{\frac{(\sigma_x^2(\tau_s, \omega))^2}{(\sigma_b^2(\tau_s, \omega))^2} + \frac{\sigma_x^2(\tau_s, \omega)}{\sigma_b^2(\tau_s, \omega)}}{\frac{(\sigma_x^2(\tau_s, \omega))^2}{(\sigma_b^2(\tau_s, \omega))^2} + 1 + 4\frac{\sigma_x^2(\tau_s, \omega)}{\sigma_b^2(\tau_s, \omega)}}} \quad (4.25)$$

Using (4.4), the final noise suppression function $V(\tau_s, \omega)$ that minimizes (4.25) is given by

$$V(\tau_s, \omega) = \sqrt{\frac{\gamma^2(\tau_s, \omega) + \gamma(\tau_s, \omega)}{\gamma^2(\tau_s, \omega) + 4\gamma(\tau_s, \omega) + 1}} \quad (4.26)$$

where $\gamma(\tau_s, \omega)$ is an a priori SNR defined as follows:

$$\gamma(\tau_s, \omega) = \frac{E[X^2(\tau_s, \omega)]}{E[B^2(\tau_s, \omega)]} = \frac{\sigma_x^2(\tau_s, \omega)}{\sigma_b^2(\tau_s, \omega)} \quad (4.27)$$

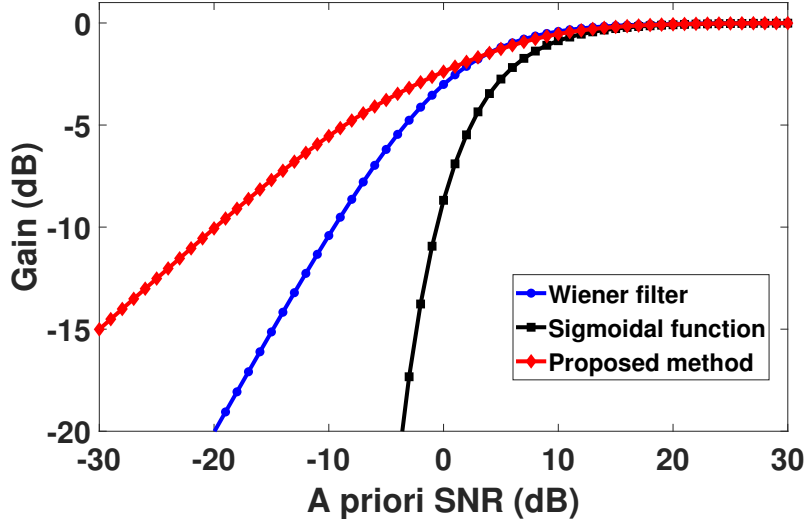


Figure 4.2: Comparison of the noise suppression functions with different a priori SNRs

Comparing $V(\tau_s, \omega)$ in (4.26) with Wiener function $W(\tau_s, \omega)$ in (4.3) and sigmoidal function $g(\tau_s, \omega)$ in (4.6) and observing the same in Fig. 4.2, the proposed method (PM) provides better gain for lower SNR levels than the Wiener filter and sigmoidal function. It is evident that the PM is the most allowable (less aggressive) suppression function.

4.3 Processing Steps for proposed method

This study sampled the noisy speech signal with a sampling rate of 44100 Hz and transformed it into a frequency domain with the FFT of two times the window length, considering a 20ms frame length. Additionally, a frame-shift of 12 ms was applied. The extracted speech was windowed and transferred to FFT for getting spectral analysis. The absolute value of the spectral bands served as the magnitude. Similarly, the phase was extracted for reconstructing the original signal. An a priori SNR ($\gamma(\tau_s, \omega)$) was calculated using a voice activity detector and a decision direct method (Martin 2001) (Ephraim and Malah 1984), and was used as the basis for evaluating the proposed and WF functions. In some studies (Hu and Loizou 2010) (Chiea *et al.* 2021), $\gamma(\tau_s, \omega)$ was calculated from the available clean and noise samples individually. However, in this case, $\gamma(\tau_s, \omega)$ is calculated from the noisy-speech spectrum for practical purposes as shown in Fig. 4.3. This study computed the enhanced speech from the proposed method along with the WF.

Verifying the performance of the PM on actual cochlear implants can be complicated by various factors such as the availability of neural survival, duration of the deafness, insertion depth, etc (Kan *et al.* 2013). The above factors can confound the outcome, so it would be better to test it with acoustic simulation before testing it on actual CI users. If simulation results are positive, the algorithm can be tested on actual CIs as well. Vocoders are commonly used to replicate some of the characteristics of CI signal processing in CI research (Loizou 2006).

This study processed the estimated speech through a 16-channel sinewave vocoder (CI simulator) as shown in Fig. 4.3. The frequency range of the channels was selected using a gamma-tone filter bank with a range of 80 to 7562 Hz (Ngamkham *et al.* 2010) (Poluboina *et al.* 2022).

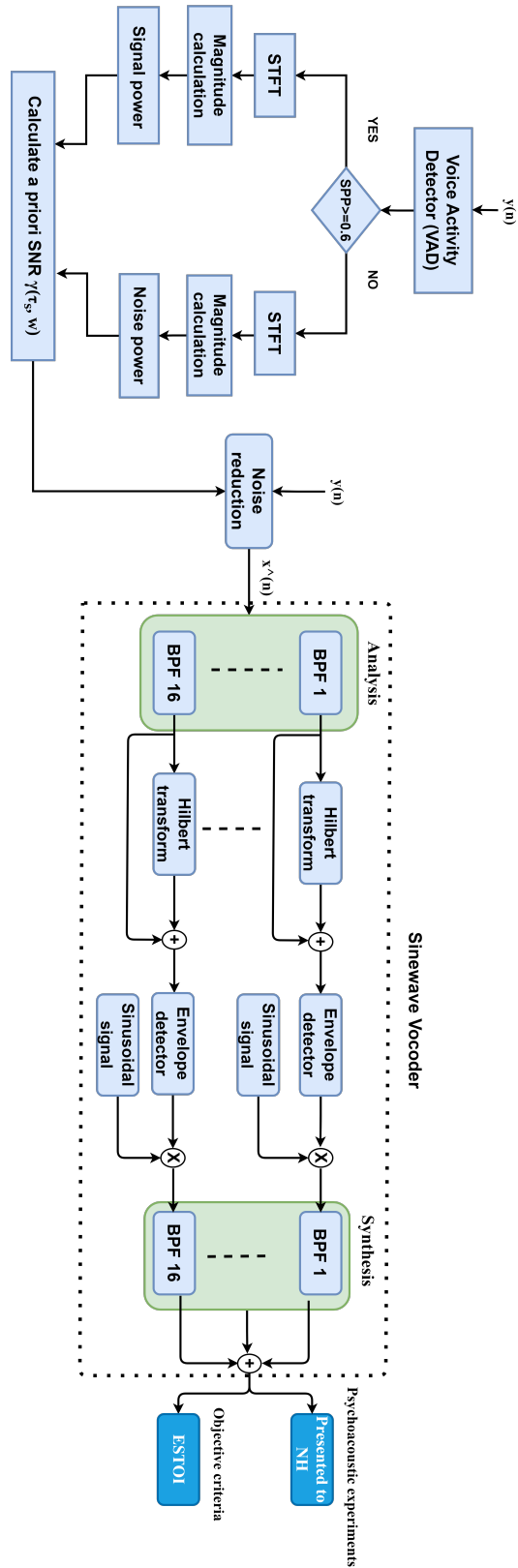


Figure 4.3: Block diagram of steps involved in psychoacoustic studies and objective assessment.

4.4 Simulation Results

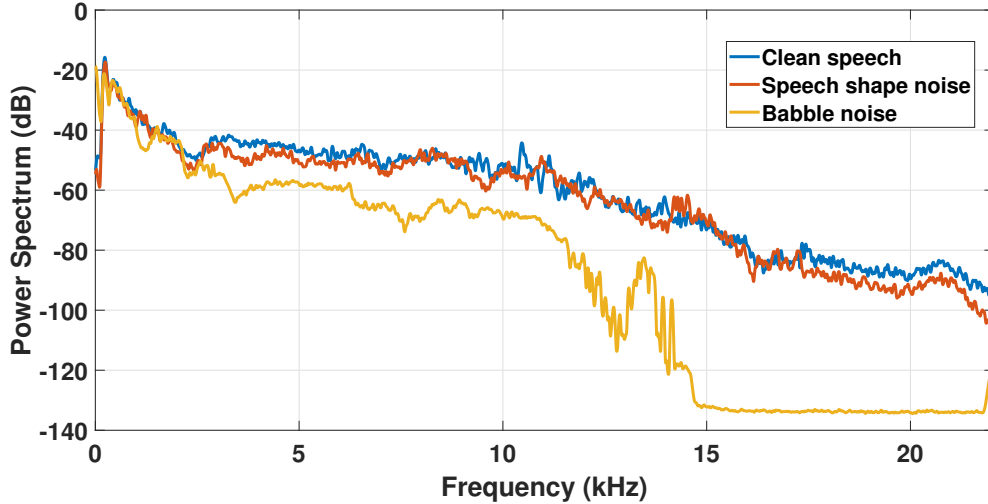


Figure 4.4: Power spectrum of speech shape noise and 4-talker babble noise.

To estimate the performance of the PM, the study conducted both perceptual and acoustic evaluations for speech in noise. This study selected two different noises here: speech shape noises and 4-talker babble noise, and their power spectrum was compared with clean speech spectra as shown in Fig. 4.4. Two different noises, 4-talker babble noise and speech shape noise are added to clean speech (Avinash *et al.* 2010) at different SNRs (15, 10, 05, 0, -5, -10, -15) in dB. Fig. 4.5 shows the waveform and the spectrogram of clean speech, noisy speech (noise at -5 dB), and noisy speech modified by the PM, WF, and sigmoidal function. Here, the magnitude of a (noisy/processed) speech envelope at every time interval is related to the intensity. The spectrogram obtained by the WF (Fig. 4.5(i)) and PM (Fig. 4.5(h)) seems to have improved the signal strength compared with the spectrogram obtained using the sigmoidal function (Fig. 4.5(j)) and the unprocessed (Fig. 4.5(g)).

The Fig. 4.6 shows the waveform of a small segment of clean speech (Fig. 4.6(a)), noisy speech (noise added at -5 dB) (Fig. 4.6(b)), and noisy speech processed by the PM (Fig. 4.6(c)), WF (Fig. 4.6(d)), and sigmoidal functions (Fig. 4.6(e)). At the red ellipse time slot, the proposed method waveform shows the speech component and small noise, whereas the WF method completely vanishes the speech component. Hence, the proposed method provides better gain than the other methods, predominantly at negative SNRs.

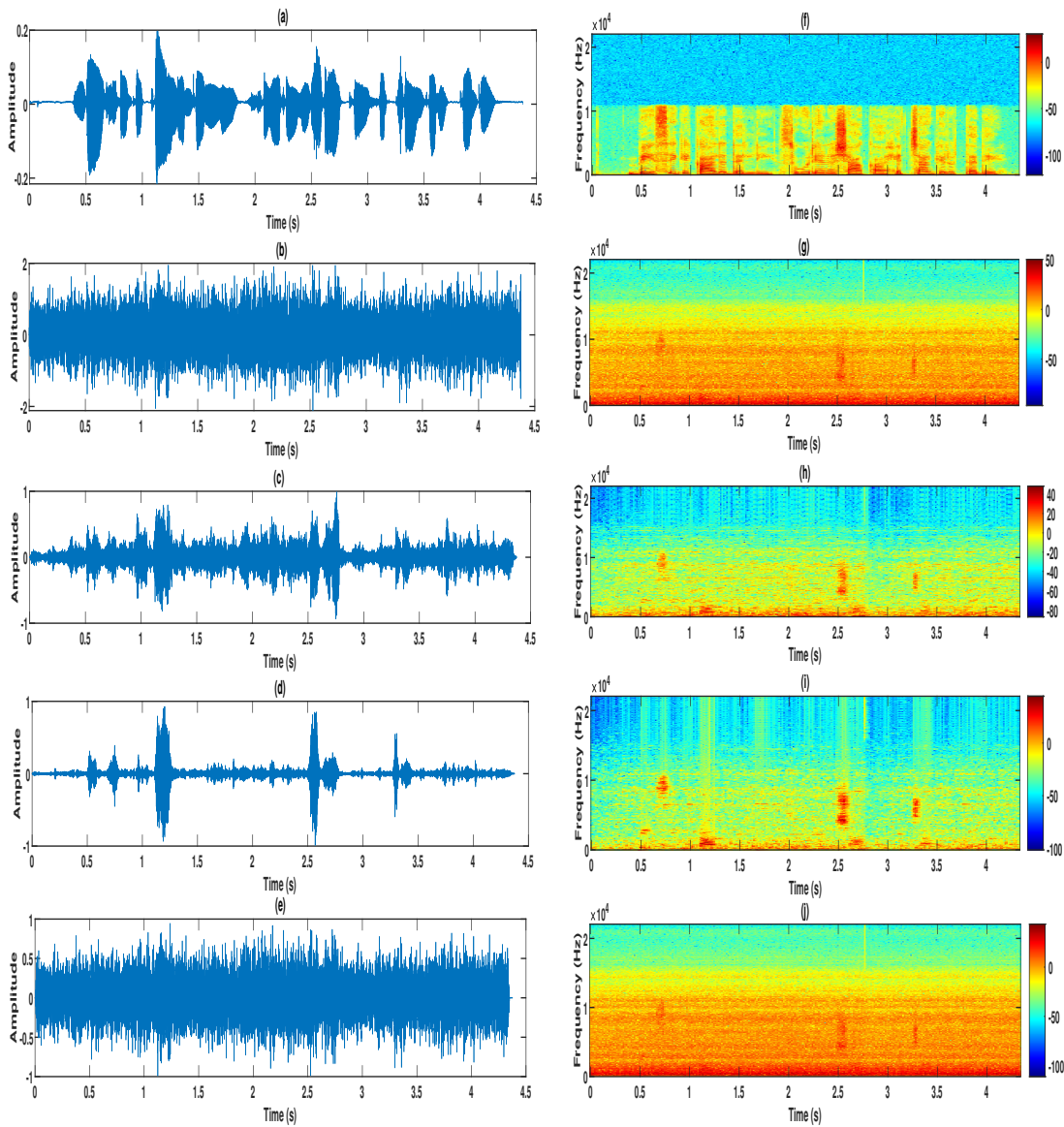


Figure 4.5: The waveform of (a) clean, (b) noisy, speech signals enhanced by (c) PM, (d) WF, and (e) Sigmoidal function. Spectrogram representation of (f) clean, (g) noisy, and speech signals enhanced by (h) PM, (i) WF, and (j) Sigmoidal function.

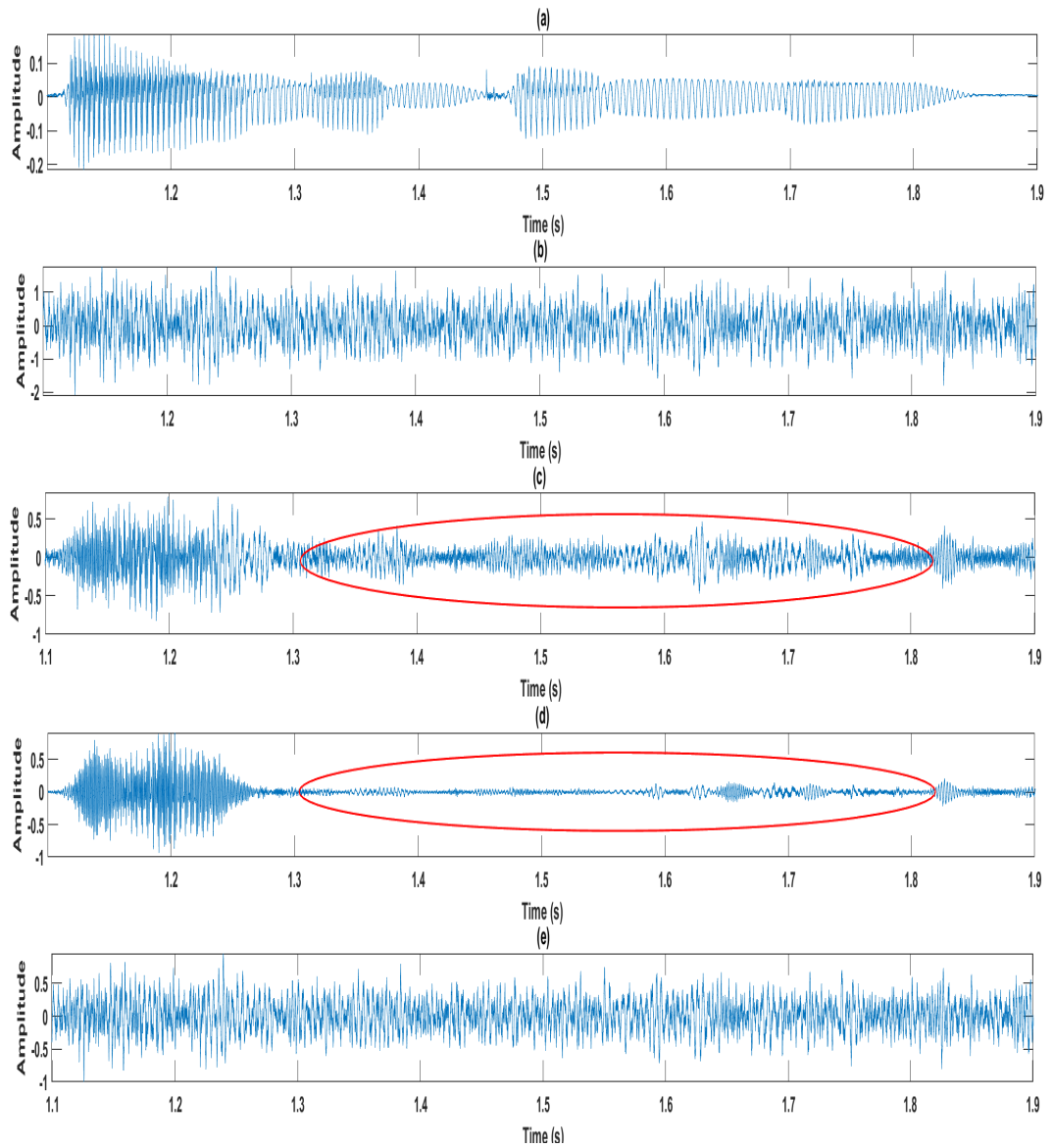


Figure 4.6: The expanded waveform of (a) clean, (b) noisy, speech signals enhanced by (c) PM, (d) WF, and (e) Sigmoidal function.

4.4.1 Objective evaluation of the noise suppressed functions using MSE

The proposed method was compared with the WF and sigmoidal function based on the mean square error between the clean signal and estimated signal envelopes.

$$MSE_{env}(s, c) = \frac{1}{T} \sum_{n=1}^T \left[|X_a(n)| - |\hat{X}_a(n)| \right]^2 \quad (4.28)$$

and the MSE between desired and estimated signal

$$MSE_{signal}(s, c) = \frac{1}{T} \sum_{n=1}^T \left[X(n) - \hat{X}(n) \right]^2 \quad (4.29)$$

Where T represents the total number of samples present in each noisy speech, and s represents the total number of noisy sentences.

Table 4.1: Mean square error at different SNR levels with speech shape noise

Method	Mean square error						
	+15 dB	+10 dB	+5 dB	0 dB	-5 dB	-10 dB	-15 dB
PM	0.0012	0.00135	0.0013	0.0014	0.0015	0.0016	0.0017
WF	0.0013	0.0014	0.0014	0.0017	0.0018	0.0019	0.002

Considering $c = 16$ channels and $s = 49$, a total of 784 samples were used for evaluating both (4.28) and (4.29). Table 4.1 displays the mean square error values at various SNR levels. The evaluated mean square errors shown in Fig. 4.7 indicate that at different SNR levels, the PM (red) offers lower MSE values than WF (blue). Moreover, this study observes that the relative effectiveness of the proposed method for predicting the speech envelope (as compared to the WF) has increased when the SNR decreases. These results support the theory presented in the previous sections 4.2. Since the WF attempts to reduce the MSE between clean and its estimated signal, the proposed method can be used to estimate the MSE between the clean and the estimated envelopes.

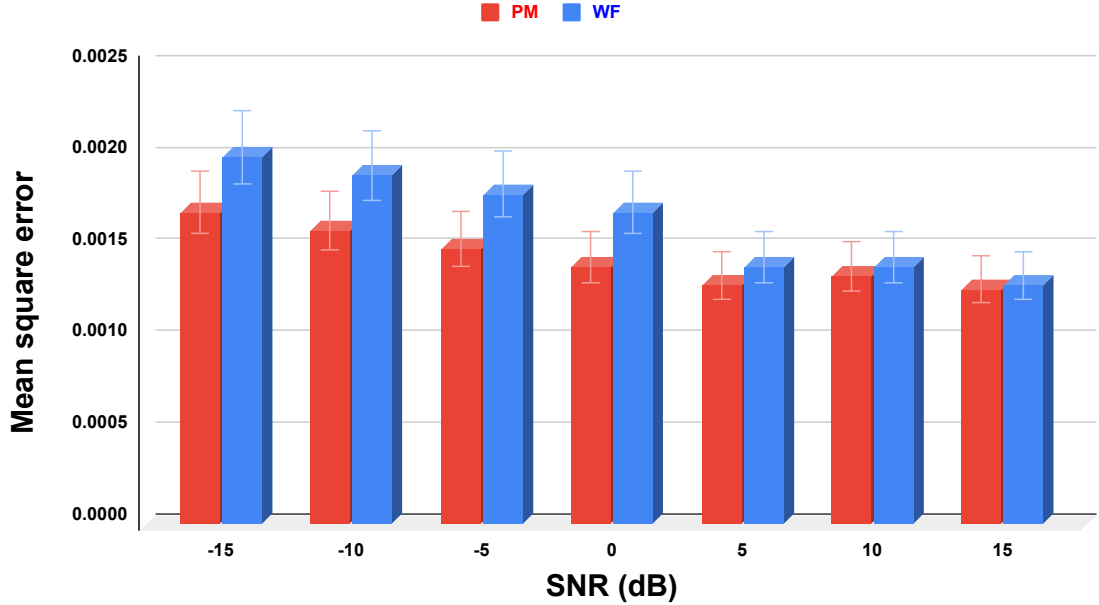


Figure 4.7: WF (blue) and PM (red) represent the mean square error concerning input SNR.

4.4.2 Speech intelligibility of cochlear implants

This study evaluated the performance of the proposed method on CIs objectively, using speech-to-reverberation modulation energy ratio (SRMR). Specifically, the SRMR-CI metric evaluates the CI signal processing speech intelligibility (Santos and Falk 2014). The SRMR-CI metric is computed over a four-stage process. In the first stage, the input signal $\hat{x}(n)$ passes through a 16-channel gammatone filter bank or a 22-channel filter bank that matches the one used in CI devices to simulate cochlear processing. In the second stage, temporal envelopes are estimated using Hilbert transform for each channel $e_c(n)$, where c represents number of channels. The discrete Fourier transform is then used after windowing the temporal envelopes (256 ms frames with 64 ms frame-shifts) to create envelope frame (with the frame index m). For each channel at each frame, the modulation spectral energy must be determined:

$$E_c(m, k) = |FT(e_c(m, k))|^2 \quad (4.30)$$

where k represents the modulation frequency bin. In the third stage, the modulation

frequency bins are arranged into eight overlapping bands with centre frequencies logarithmically separated between 4 and 64 Hz for CI users and 4 and 128 Hz for NH listeners (Santos *et al.* 2013). This simulates frequency selectivity in the modulation domain (Ewert and Dau 2000). Finally, the SRMR value is determined by the ratio of the mean energy content of the first four modulation bands (between 3 to 20 Hz) to the mean energy content of the last four modulation bands (between 20 to 160 Hz).

Table 4.2: Speech intelligibility of CIs according to SRMR-CI metrics with different SNR levels.

SNR in dB	Noisy	Sigmoidal function	WF	PM
15	0.80	0.83	0.89	0.895
10	0.574	0.584	0.89	0.894
5	0.30	0.34	0.70	0.682
0	0.134	0.14	0.54	0.532
-5	0.08	0.08	0.43	0.425
-10	0.07	0.072	0.35	0.385
-15	0.067	0.071	0.24	0.35
-20	0.063	0.071	0.213	0.34
-25	0.06	0.071	0.21	0.34

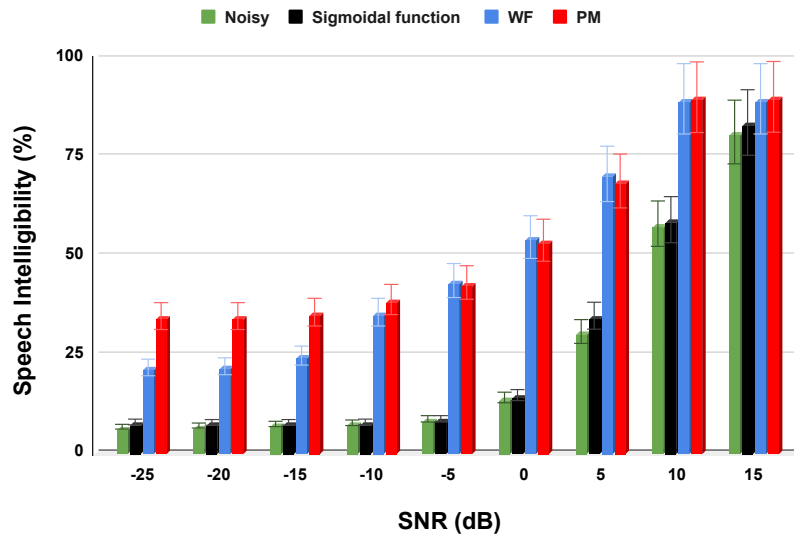


Figure 4.8: Speech intelligibility of CIs according to SRMR-CI metrics with different SNR levels.

Table 4.2 displays the evaluation of SRMR metrics across various SNR levels. The proposed method consistently provides higher scores across all SNR levels. Fig. 4.8 shows that speech intelligibility decreases with unprocessed speech, WF, and sigmoidal functions at low SNR levels, particularly $SNR < -5dB$ compared to PM. The PM shows results similar to the Wiener filter at positive SNR levels.

4.4.3 Acoustic assessment of speech in noise

The Extended Short-Time Objective Intelligibility (ESTOI) metric was applied to assess the objective evaluation of speech intelligibility in noise, based on the correlation between processed noisy speech and clean speech. ESTOI is highly relevant to human speech intelligibility, according to previous studies (Mesnildrey *et al.* 2016). ESTOI values are evaluated in terms of speech intelligibility index values (D) at various SNRs. The D values range from 0 to 1, and higher values suggest better speech intelligibility. An ESTOI score is computed in three steps: (1) Each subband’s temporal envelope is obtained after passing the signals through a filter bank of one-third octave; (2) the distance between the clean speech and processed speech short-time envelope spectrograms is estimated after time and frequency normalization, resulting in intermediate indices for short-time intelligibility; (3) the final intelligibility index D is derived by averaging the intermediate indices. More information on the three steps of the ESTOI measurement is provided in (Mesnildrey *et al.* 2016).

(a) Objective evaluation of speech intelligibility: The speech intelligibility index (D) values were computed for six different lists with two noise conditions (speech shape noise and babble noise) for three noise reduction strategies at different SNRs (+15 dB, +10 dB, +05 dB, 0, -5 dB, -10 dB, -15 dB). A general trend was observed by analyzing the D values. The D values decreased when the SNR decreased from +15 dB to -15 dB in the three methods. Fig. 4.9 shows the average ESTOI scores at seven different SNR levels for the speech shape noise.

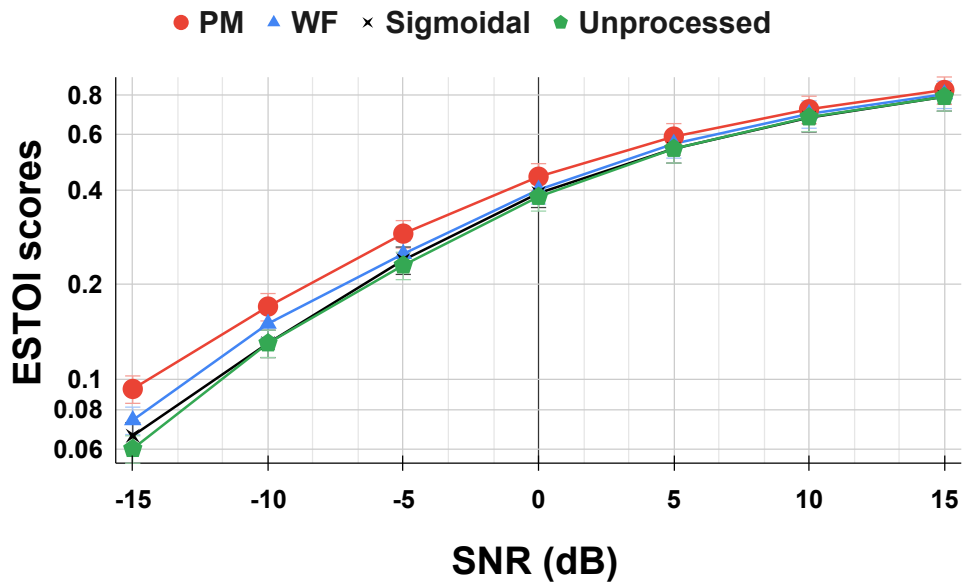


Figure 4.9: ESTOI scores for speech signals corrupted by speech shape noise at different SNR levels.

Table 4.3: ESTOI values (D) for each noise reduction method with speech shape noise

SNR in dB	Unprocessed	Sigmoidal function	WF	PM
15	0.79	0.79	0.805	0.83
10	0.68	0.678	0.69	0.721
5	0.54	0.54	0.56	0.59
0	0.38	0.39	0.40	0.44
-5	0.23	0.239	0.25	0.29
-10	0.13	0.13	0.15	0.17
-15	0.06	0.066	0.074	0.093

Table 4.3 gives the average ESTOI scores (D) at seven different SNR conditions with the speech shape noise. The proposed method obtained maximum D values for all SNR levels in speech shape noise compared to the WF and sigmoid function.

Table 4.4: ESTOI values (D) for each noise reduction method with babble noise

SNR in dB	Unprocessed	Sigmoidal function	WF	PM
15	0.81	0.81	0.814	0.815
10	0.71	0.71	0.73	0.728
5	0.59	0.59	0.618	0.613
0	0.45	0.45	0.44	0.47
-5	0.30	0.3	0.27	0.31
-10	0.184	0.18	0.142	0.173
-15	0.087	0.085	0.053	0.079

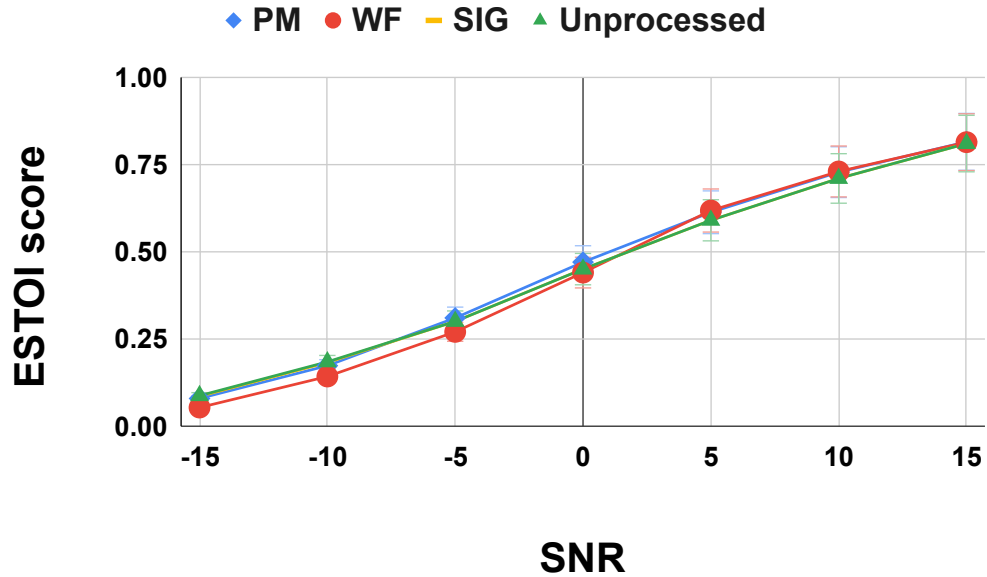


Figure 4.10: ESTOI scores for speech signals corrupted by babble noise at different SNR levels.

Similarly, the proposed method's speech intelligibility (D) values were nearly identical to the other methods when dealing with babble noise as shown in Table 4.4. The average ESTOI scores at seven different SNR levels with 4-talker babble noise are displayed in Fig. 4.10. It can be observed that the PM (red) exhibits comparable performance to both WF and sigmoidal functions.

4.4.4 Subjective evaluations

The output speech stimuli of the CI simulator were presented to the normal hearing (NH) volunteers through headphones at the most comfortable level (40 dB speech recognition threshold (SRT)) for conducting a psychoacoustic test. Similarly, ESTOI was used to determine the speech intelligibility of the estimated signal and target signal in acoustic evaluation tests.

(a) Participants: This study included six NH individuals with no previous complaints of hearing problems. The sample size used in the present study fulfills the minimum required sample size for psycho-physical research (Anderson and Vingrys 2001). The participants were 25 years old on average (with a 3.4-year standard deviation) as shown in Table 4.5. The individuals have given written permission before participating in this study, by following the Helsinki Declaration. The local Ethics Committee has given its approval to the study (Approval Number: NITK/EC/Ph.D/284/2021).

Table 4.5: Normal hearing Participants details.

Participant	Age	Gender
NH1	21	Male
NH2	29	Female
NH3	25	Male
NH4	29	Female
NH5	22	Male
NH6	24	Female

(b) Dataset and Stimuli presentation: For this experiment, the given input signal is a noisy speech signal. Two different noises, 4-talker babble noise and speech shape noise are added to clean speech at different SNRs (15, 10, 05, 0, -5, -10, -15) in dB. The pre-processed signals in MATLAB were given to NH volunteers via Sennheiser HD280pro headphones. A practice trial has been given to all participants to avoid potential learning effects. Once the individuals had become familiar with the task, they were subjected to the actual perceptual test. The sentence list for each signal processing condition was randomized for each participant. The testing sequence was also randomly assigned to each participant. The speech recognition test in noise had 7 different lists (Avinash *et al.* 2010), with each list having 7 sentences, and each sentence having 5 keywords. These clean speech sentences have information up to

10kHz. The standardized QuickSIN protocol (Avinash *et al.* 2010) wherein, the first sentence of every list begins at +15 dB SNR, and the remaining sentences were given with decreasing SNR by +5dB sequentially. Participants were required to identify the words in the sentences as they hear. The responses were collected in written form for further evaluation.

(c) Perceptual evaluation: This study calculated the total number of keywords identified correctly by every participant in each method. The mean speech recognition scores were calculated, representing the participants’ average speech intelligibility (proportion correct) at the corresponding SNR. The speech recognition threshold in noise (SRTN) was calculated using Finney’s (1952) Spearman Karber Equation given by:

$$SRTN = i + (d/2) - (d * correct/W) \tag{4.31}$$

Where i = initial presenting SNR

d = step size (+5dB)

W = identified keywords per decrement with SNR

correct = number of key words correctly identified

Perceptual measure with speech shape noise: The mean speech recognition scores were evaluated for different noise reduction methods at different SNR levels with speech shape noise, as shown in Table 4.6.

Table 4.6: Mean speech intelligibility (PC) of volunteers with speech shape noise

Method	Mean speech intelligibility score						
	+15 dB	+10 dB	+5 dB	0 dB	-5 dB	-10 dB	-15 dB
Proposed method	1	1	0.966	0.933	0.2	0	0
Wiener filter	0.966	0.966	0.866	0.633	0.03	0	0
Sigmoidal	1	0.966	0.866	0.633	0	0	0
Unprocessed	0.866	0.966	0.933	0.166	0	0	0

The proportion correct score for three noise reduction methods was plotted on the Gaussian psychometric function as shown in Fig. 4.11.

As shown in Fig. 4.11, the PM (red) was significantly more effective in providing better speech intelligibility compared to the WF (blue) and sigmoidal function (black), especially at $SNR \leq 0$.

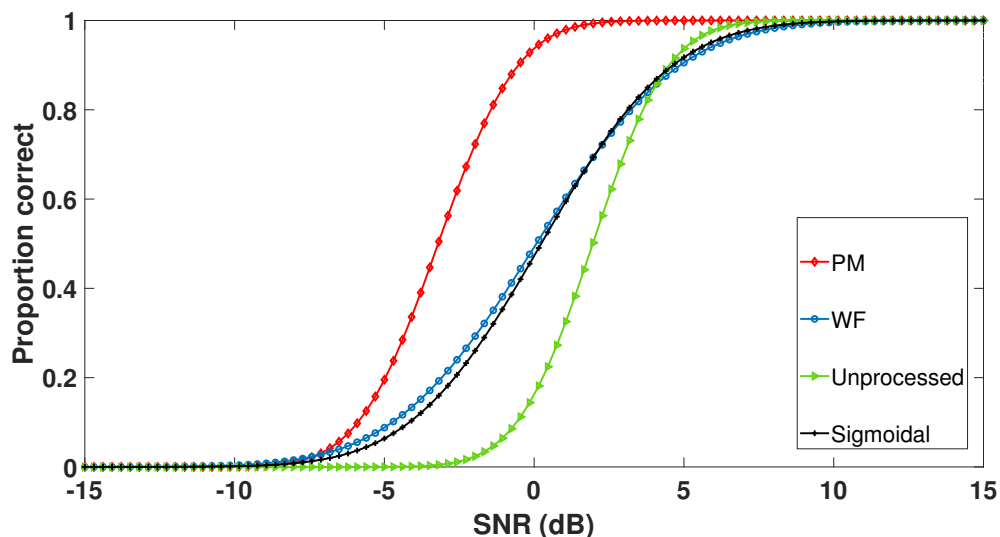


Figure 4.11: The psychometric plots for the speech recognition ability of volunteers with speech shape noise.

Table 4.7: The average SRTN of each noise-reduction method with speech shape noise

Method	SRTN in dB
Proposed method	-3.167
Wiener filter	0.167
Sigmoidal function	0.23
Unprocessed	2.833

The SRTN was calculated for each method as shown in Table 4.7. Hence, Table 4.7 indicates that the SRTN of the PM is very low compared to WF and sigmoidal functions. Evidence shows that PM requires minimum SNR for at least 50% speech perception.

Statistical analysis:

SRTN was evaluated using the Spearman-Karber equation (Finney, 1952) on each participant, as shown in Table 4.8. The calculated SRTN goes through statistical analysis, and one-way ANOVA with repeated measurements was used to investigate the noise reduction effect. The F-statistic from a repeated measures ANOVA is reported as:

$$F(df_{between}, df_{within}) = F_{value}, p = p_{value} \quad (4.32)$$

where df is degrees of freedom between the methods and within the methods. If the

Table 4.8: SRTN with speech shape noise

Participant	SRTN in dB			
	PM	WF	Sigmoidal	Unprocessed
NH1	-7.5	2.5	2.5	2.5
NH2	-2.5	-1.5	-1.5	2.5
NH3	-3.5	-3.5	-2.5	3.5
NH4	-1.5	2.5	1.5	1.5
NH5	-2.5	0.5	0.5	4.5
NH6	-1.5	0.5	0.5	2.5

p-value is 0.05, the F value has a 5% chance of being incorrect and a 95% chance of being true. The hypothesis test is significant statistically if the p value is lower than the significance level (0.05). For pairwise comparisons, a series of one-tailed paired ' t ' tests with the alternate hypothesis of, "measure 1 is less than measure 2" was performed. Since ANOVA revealed a significant difference, this study conducted pairwise comparisons.

Table 4.9: Quality assessment based on statistical tests of psychoacoustic experiments

Measure 1	vs.	Measure 2	t	p
Proposed method	vs.	Unprocessed	-5.809	0.001
Wiener filter	vs.	Unprocessed	-2.219	0.039
Sigmoidal function	vs.	Unprocessed	-2.697	0.021
Proposed method	vs.	Sigmoidal function	-2.411	0.03
Proposed method	vs.	Wiener filter	-2.294	0.035

Table 4.9 depicts the variables representing measure 1 and measure 2 for the comparisons as well as the corresponding ' t ' & ' p ' values. The value of ' t ' is determined by calculating and expressing the difference in terms of standard error units. When the ' t ' value is high, it indicates more substantial evidence against the null hypothesis. A point noteworthy of statistical inference is that the smaller the SRTN value better is the performance. Noise reduction had a significant main effect on speech recognition in the presence of speech-shaped noise ($p < 0.05$). The analysis revealed that the SRTN of the PM is significantly better than without noise reduction (Unprocessed) with t (t -distribution) = -5.809, $p=0.001$. In addition, the SRTN with the WF is con-

siderably better than the Unprocessed with $t = -2.219$, $p = 0.039$. Hence, the analysis revealed that both PM and WF have significantly improved speech recognition with speech shape noise as compared to that without noise reduction. However, the proposed method has shown significant improvement in SRTN results compared to the traditional wiener filter with $t = -2.294$, $p = 0.035$.

Perceptual measure with babble noise:

Table 4.10: Mean speech intelligibility (PC) of volunteers with babble noise

Method	Mean opinion score						
	+15 dB	+10 dB	+5 dB	0 dB	-5 dB	-10 dB	-15 dB
Proposed method	0.966	1	0.533	0.266	0	0	0
Wiener filter	0.966	0.766	0.4	0.166	0	0	0
Sigmoidal	0.966	0.766	0.533	0.2	0	0	0
Unprocessed	0.966	0.866	0.533	0.166	0	0	0

A masker of four-talker Kannada language babble was selected. This study intends to reflect the properties of actual listening. Further, babble efficiently reduces the intensity of amplitude modulation of speech over steady spectrum noise. The mean speech recognition scores were evaluated for different noise reduction methods at different SNR levels with babble noise, as shown in Table 4.10.

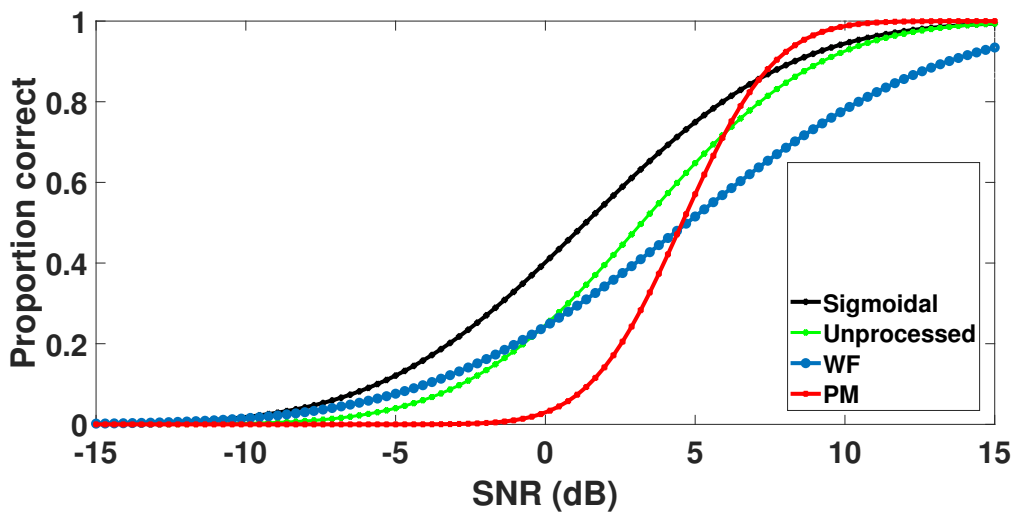


Figure 4.12: The psychometric plots for the speech recognition ability of volunteers with babble noise.

Table 4.11: The average SRTN of each noise-reduction method with 4 talker babble noise

Method	SRTN in dB
Proposed method	3.667
Wiener filter	4.167
Sigmoidal function	4.16
Unprocessed	3.833

The proportion correct score for three noise reduction methods was plotted on the Gaussian psychometric function as shown in Fig. 4.12. The proposed method provides comparable speech intelligibility to the other techniques with babble noise, and when compared to speech shape noise, the speech recognition scores of all methods with babble noise are provided significantly lower scores. The determined SRTN underwent statistical analysis, and the noise reduction effect was measured using a one-way ANOVA with repeated measurements. There was no significant improvement of noise reduction block on speech recognition score with babble noise ($p=0.893$). However, the observation of mean values indicates that the SRTN with the proposed method is slightly better than the traditional WF results shown in Table 4.11 and Fig. 4.12. Hence, compared to WF and Unprocessed, PM offers a marginal improvement in speech recognition when speech babble is present.

4.5 Discussion

Compared to the WF time-frequency mask, the proposed algorithm implementation requires extra 3 multiplications, 2 additions, and 1 square root calculation. This is due to the fact that the proposed time-frequency mask was derived by minimizing the MSE between the squared envelopes of the enhanced speech and its clean speech. The complexity of the proposed algorithm can be understood by comparing its equation (4.26) with that of the WF (4.3). However, the proposed method outperforms the WF method in terms of speech intelligibility, as shown in Table 4.7, with comparable complexity.

As seen in Fig. 4.2, the noise suppression of the PM is softer (most allowable) than that of the WF method. CIs requires more aggressive noise suppression (Mauger *et al.* 2012) (Serizel *et al.* 2014). However, aggressive noise suppression should only be imple-

mented if it preserves speech components. Any implementation of an aggressive WF method would compromise the speech component as well, thereby reducing speech intelligibility. The listeners with CI rely on the envelope for speech intelligibility. Therefore the preservation of temporal envelope is essential for speech intelligibility. Aggressive noise reduction (Chung 2007), might have a negative impact on the envelope. Earlier studies also have shown that less aggressive noise reduction resulted in better speech intelligibility than the more aggressive gain suppression function (Chiea *et al.* 2021)(Mourao *et al.* 2020)(Loizou 2013). This supports our findings wherein the gain function derived in the current study is less aggressive than the WF as well. Therefore, a good noise reduction algorithm should provide an optimum trade-off between the magnitude of noise suppression and preservation of speech cues.

Perceptual data indicated that the speech recognition scores improved significantly with WF noise suppression, especially with PM. The pairwise comparisons revealed that the speech recognition scores were significantly better with PM than WF and sigmoidal function. The sigmoidal function does not work for negative SNR levels, which is a crucial region for people with hearing impairments, as shown in the noise suppression function and the perceptual data. It is well-known that the traditional WF can suppress noise, but due to its aggressive nature, some of the speech’s spectral content is also lost during noise suppression, as shown in Fig. 4.5. Comparison of clean speech (a) and Enhanced Speech using WF (d) of Fig. 4.5 reveals the loss of target speech component which would have negatively affected the speech intelligibility. On the other hand, the PM provides an optimum trade-off between noise suppression and speech intelligibility. Hence, PM suppresses noise while preserving important cues for speech intelligibility, resulting in a better speech recognition score than the other two methods. According to the mean data (Table 4.7), the proposed method improved the speech recognition threshold in noise (SRTN) by 6 dB SNR in comparison to the unprocessed (noisy) data. This may lead to an improvement in speech recognition by almost 60% in real-life situations (Venema 1999).

A similar analysis was done with speech babble noise. However, the statistical analysis and subjective analysis indicated that all three methods do not improve the SNR because all three methods have been implemented in conjunction with the voice activity detector. The background noise itself is speech, and the algorithm detected the noise segment based on the voice activity. However, the speech babble is voice-based, so the algorithm fails to distinguish target and mask signals. In real-time applications,

most of the algorithms fail when background noise is speech itself. In such a scenario, it is ideal to provide more cues to the auditory system to segregate target speech and noise like temporal fine structures. Hence, in one of the previous chapters, this work proposed how effectively TFS cues can be encoded to improve speech recognition in noise, so the study recommends these MSE minimization methods for improving SNR in non-speech noise scenarios. However, for speech noise situations there is a need to investigate another method or provide more cues for segregating speech and noise.

4.6 Summary

In this chapter, a novel noise reduction technique has been proposed for cochlear implants and its performance was compared with traditional WF and sigmoidal functions. Overall perceptual and objective analyses indicated that PM is more efficient in improving speech intelligibility when compared to sigmoidal function and traditional WF. Thus, the proposed noise reduction technique has potential implication in CI and a further study can be conducted on actual CI users.

Chapter 5

DEEP DENOISING METHOD FOR IMPROVING SPEECH RECOGNITION

This chapter explains a novel training approach based on a Deep learning technique called Noisy2Noisy_{avg} (N2N_{avg}) for speech enhancement and denoising. The mathematical derivation was given to prove the advent of using the N2N_{avg} strategy over the Noise2Noise (N2N) strategy. The target and input of a DCU-Net were trained using only noisy speech samples. The proposed method was compared with the traditionally available speech-denoising methods.

5.1 Introduction

One of the most critical things of the speech enhancement scheme is to separate clean speech from noise, given the speech and noise mixture input. Recent developments in deep learning have improved the speech enhancement schemes thereby achieving high-performance standards. Typically, most traditional speech de-noising methods utilize supervised training methods (Su *et al.* 2020, Defossez *et al.* 2020, Fu *et al.* 2021, Wang *et al.* 2021) such as Noise-to-Clean (N2C). In N2C methods as shown in Table 5.1, networks use noisy audio signals as the training input and perfectly clean audio signals as the target. Nevertheless, the performance of deep learning methods depend on the number of clean data sets available for training purposes. Obtaining clean data sets with many samples is a challenge, especially in languages with low resources. The

availability of the resources depend on various factors including the cost associated with soundproofing and high-precision sound recording equipment.

Table 5.1: Literature review of training methods and Models used for speech enhancement.

Method	Model	Clean dataset	Noise
N2C	Segan (Pascual <i>et al.</i> 2017)	Voice Bank	2 Synthetic and 8 from the DEMAND
	CRNN (Zhao <i>et al.</i> 2018)	Synthetic	25 different types of noise
	DCUnet (Choi <i>et al.</i> 2019)	Voice Bank	DEMAND
	Segan (Baby and Verhulst 2019)	Voice Bank	2 Synthetic and 8 from the DEMAND
	Demucs (Defossez <i>et al.</i> 2020)	LIBRISPEECH	DNS dataset
	CAUnet (Wang <i>et al.</i> 2021)	Voice Bank	2 Synthetic and 8 from the DEMAND
N2N	N ₁ 2N ₂ (Kashyap <i>et al.</i> 2021)	Voice Bank	UrbanSound8K

To overcome the limitations of the lack of clean data set, Lehtinen et al attempted to denoise and reconstruct the image using deep learning models without a clean image as the target (Lehtinen *et al.* 2018). The authors demonstrated that, under certain circumstances, images could be reconstructed using just noisy images as a reference and they termed it as 'Noise2Noise (N2N) strategy'. Later, this N2N strategy was extended to audio signal processing as well. To get around the limitation of the lack of a clean speech data set, few studies have experimented with training without a clean speech (Kashyap *et al.* 2021)(Alamdari *et al.* 2021). Both studies demonstrated the possibility of training convolutional neural networks to denoise speech without using clean speech samples. Despite the N2N strategy, the authors recommended keeping the loss function (between estimated and target speech) as small as possible to achieve better results. In the derivation of the loss function in (Kashyap *et al.* 2021), it has been observed that the variance of noise distribution is directly proportional to the loss and inversely proportional to the sample datasize. Thus to reduce the loss, the sample data size should be large enough (Xu *et al.* 2013) which increases the computation time (Lu *et al.* 2013). As a necessary consequence, a novel technique is needed to train the deep-learning network without using clean data while still limiting the size of the training dataset.

The standard speech-denoising networks used in real-time applications concentrate on spectrogram magnitude estimation of the enhanced speech and phase of the noisy speech (Xu *et al.* 2014)(Nugraha *et al.*, 2016)(Graiss and Plumbley 2017)(Takahashi

et al. 2018). However, both the magnitude and the phase of the enhanced signal are essential for optimal reconstruction. Hence, this study uses Deep complex U-net (DCU-net) (Choi *et al.* 2019) where both the magnitude and phase of the complex masks are used. This study has proposed a novel training method where both the input and the target of the 10-layered DCU-net are noisy data. Training at the target use noisy speech data wherein the noise was an average of two independent noises. Training at the input was performed using noisy speech data whose noise was uncorrelated with the noise used for training at the target. This methodology helps to overcome the strong dependency on the data size seen in the N2N (Kashyap *et al.* 2021) method by minimizing the variance with the use of average of noises. The performance of the suggested training method was evaluated using various objective speech intelligibility metrics at different SNR levels.

Our contributions in this work are as follows: (1) Deriving a mathematical relation to prove the advent of using the Noisy2Noisy_{avg} (N2N_{avg}) strategy, (2) Bringing forward a novel training approach, (3) Conducting an extensive experiment to examine various training conditions, and (4) Evaluating the results of the experiment.

The subsequent sections are organized as follows: In Section 5.2 methodology of the proposed training method is explained. Section 5.3 provides discussion of the experimental results. Section 5.4 provides summery of the chapter.

5.2 Proposed Method

Assume a Deep Learning Algorithm (DLA) with the parameters (p), loss function (L), input (y), output $f_p(y_1)$, and target (x). By resolving the optimization problem as described in equation (5.1), the DLA learns to denoise the input audio.

$$\operatorname{argmin}_p \mathbb{E}\{L(f_p(y_1), x)\} \quad (5.1)$$

This study considered two uncorrelated noises (N_1, N_2) with zero mean distribution. Two noisy speech samples, y_1 , and y_2 were generated by adding the noises $N_1 \& N_2$ individually to the clean speech x .

$$y_1 = x + N_1, \text{ and } y_2 = x + N_2 \quad (5.2)$$

The average of y_1 and y_2 can be written as the sum of the clean speech and the

average of two noises (N_1, N_2)

$$y_3 = x + \frac{N_1 + N_2}{2} \quad (5.3)$$

Clean speech samples are used as targets for training in traditional Noise2Clean (N2C) DLA techniques (Pascual *et al.* 2017, Zhao *et al.* 2018, Baby and Verhulst 2019, Azarang and Kehtarnavaz 2020, Su *et al.* 2020, Defossez *et al.* 2020, Fu *et al.* 2021, Wang *et al.* 2021). The typical deep learning approaches estimate the loss function (L_1) using the clean speech samples as targets. The optimization of L_1 is given below

$$L_1 = \operatorname{argmin}_p \mathbb{E}\{(f_p(y_1) - x)^2\} \quad (5.4)$$

In the N2N approach (Kashyap *et al.* 2021), the loss is minimized between the estimated signal and noisy signal (y_2)

$$L_2 = \operatorname{argmin}_p \mathbb{E}\{(f_p(y_1) - y_2)^2\} \quad (5.5)$$

The above equation can be rewritten in terms of L_1 and variance of N_1 according to Kashyap *et al.* 2021

$$L_2 = L_1 + \operatorname{Var}(N_1) \quad (5.6)$$

In the above equation, the variance of the noise sample distribution $\operatorname{Var}(N_1)$ is inversely proportional to the sampling size. Therefore, if the training dataset size increases, then the N2N loss (L_2) value equals the N2C loss (L_1) value. As a result, to minimize the loss, the sample data size must be sufficiently large (Xu *et al.* 2013), which increases the computation time (Lu *et al.* 2013).

In the proposed Noisy2Noisy_{avg} (N2N_{avg}) technique, the targets are not trained with clean speech samples. Instead, this study employs average noisy outputs and noisy inputs throughout the training stage. The N2N_{avg} optimization equation is defined as

$$L_3 = \operatorname{argmin}_p \mathbb{E} [(f_p(y_1) - y_3)^2] \quad (5.7)$$

The above equation can be re-written by substituting equation (5.3) in equation (5.7)

$$L_3 = \operatorname{argmin}_p \mathbb{E} \left[\left(f_p(y_1) - \left(x + \frac{(N_1 + N_2)}{2} \right) \right)^2 \right] \quad (5.8)$$

The above equation can be written in detail as follows

$$L_3 = \arg \min_p \mathbb{E}[(f_p(y_1) - x)^2 + x * (N_1 + N_2) + ((N_1 + N_2)/2)^2 - f_p(y_1)(N_1 + N_2)] \quad (5.9)$$

This study assumes that the mean values of the noises N_1 & N_2 are equal to zero and both noises are uncorrelated. Using L_1 from equation (5.4), equation (5.9) can be rewritten as:

$$L_3 = L_1 + \frac{1}{4} E[N_1^2 + N_2^2 + 2N_1N_2] \quad (5.10)$$

Since there is no correlation between the two noise distributions, the expectation of the $2N_1N_2$ term becomes zero and the mean value's square is also zero. The mean square of the noise can be expressed in terms of variance.

$$L_3 = L_1 + \frac{1}{4} [Var(N_1) + Var(N_2)] \quad (5.11)$$

Generally, the following formula is used to calculate the variance of the mean sampling distribution:

$$Var(N)_{mean} = \frac{\text{population variance}}{\text{sampling size}} \quad (5.12)$$

When the data sampling size increases, the variance of the noise distributions decreases by 4 times. Thus, the L_3 loss value reaches the L_1 value faster than the L_2 value. As a result, when the data reach a sufficient size, the loss function L_3 value can become L_1 . Therefore, without having a clean data set it is possible to achieve a better mean square error.

In the same way, if n noises (uncorrelated and have a zero mean distribution) are considered, such as $N_1, N_2, N_3, \dots, N_n$, then the average of these noises will become a noise file (N_{avg}). This study generates the following noisy target file (y_n) by combining N_{avg} with a clean (x) sample.

$$y_n = x + \frac{N_1 + N_2 + N_3 + \dots + N_n}{n} \quad (5.13)$$

Assuming that y_n is the target file, the N_2N_{avg} optimization equation (5.7) can be

written as follows.

$$L_3 = \arg \min_p \mathbb{E}[(f_p(y_1) - (x + \frac{N_1 + N_2 + N_3 + \dots + N_n}{n}))^2] \quad (5.14)$$

The previous equation can be simplified as follows (steps followed to simplify equation (5.9))

$$L_3 = L_1 + \frac{1}{n^2} [Var(N_1) + Var(N_2) + Var(N_3) + \dots + Var(N_n)] \quad (5.15)$$

This equation shows that the variance of the noise distribution is reduced by n^2 times. Thus, by considering more noisy data, the variance of noise can be reduced which in turn reduces the requirement of a large data set.

5.2.1 DCUNET Architecture

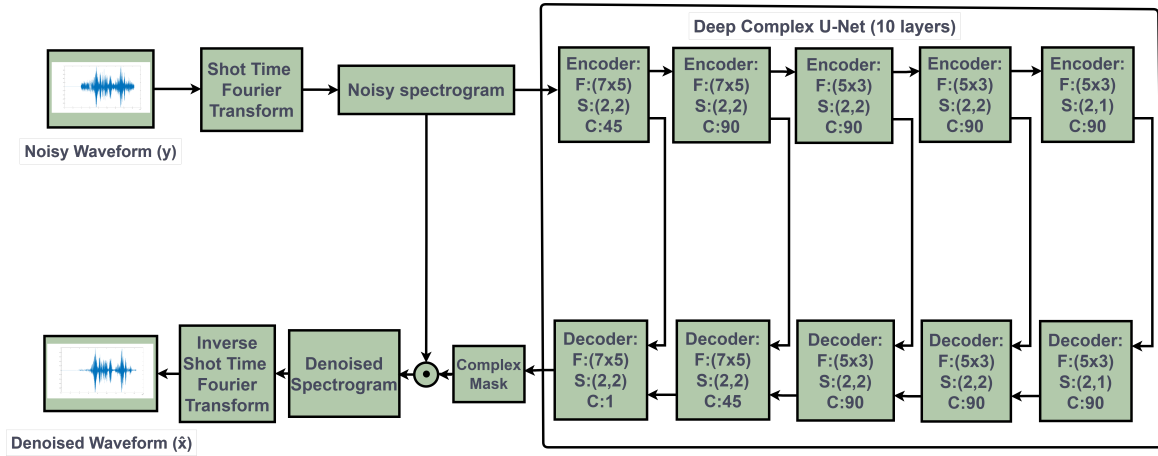


Figure 5.1: Architecture of a ten-layer DCU-net for speech denoising.

The performance of the $N2N_{avg}$ method was demonstrated using the 10-layered Deep Complex U-Net (DCU-net-10) architecture (Choi *et al.* 2019). A complex-valued masking approach was an improvement over the well-known U-Net (Ronneberger *et al.* 2015) architecture. As shown in Fig. 5.1, the noisy waveform was converted into a spectrogram $Y(t, f)$ through STFT with 16ms hop length and 64ms window size. The spectrogram can be defined as magnitude ($Y(t, f)^{mag}$) and phase ($Y(t, f)^{phase}$) components. As a result, a two-dimensional spectrogram was obtained, which can be divided into a real-valued magnitude component and a real-valued phase com-

ponent. A real-valued architecture like U-Net extracts information from magnitude spectrograms and discards some valuable information from phase spectrograms. The real-valued convolution layers cannot process complex-valued phase information. To overcome this limitation, deep complex UNET was used to process both phase and magnitude spectrogram information. Thus, the enhanced audio can be phase estimated and reconstructed more precisely.

In simplest terms, DCU-net is a complex-valued convolutional autoencoder. As described in (Trabelsi *et al.* 2017), complex convolution layers, complex weight initialization, complex batch normalization, and CReLU have been applied. There are three stages in each of the encoding and decoding processes: kernel size (F), stride sizes (S), and output channels (C) followed by batch normalization, and leaky complex ReLU (LeCReLU) as activation function. A convolutional layer with stride prevents the loss of spatial information when down sampling. When up-sampling, the dimension of the input is restored through these complex de-convolutional layers.

The mean squared error (MSE) between clean speech (x) and predicted speech (\hat{x}) on the STFT domain is a commonly used loss function for audio source separation. However, it has been found that due to the variability in the phase structure, optimizing the model using MSE in the complex STFT domain fails in phase estimation (Williamson *et al.* 2015). Since the phase information in the raw waveform is inherent, it is also possible to utilize a loss function defined in the time domain as an alternative. Hence, this work used the weighted source-to-distortion ratio loss function (lossWSDR) introduced in (Choi *et al.* 2019), which directly optimizes the popular evaluation metrics used in the time domain. The loss values range between $[-1, 1]$. The loss function is defined in terms of noisy speech (y), target (x), and estimated (\hat{x}) signals as follows:

$$Loss_{WSDR} = -\alpha \frac{\sum(x, \hat{x})}{\|x\| \|\hat{x}\|} - (1 - \alpha) \frac{\sum(z, \hat{z})}{\|z\| \|\hat{z}\|} \quad (5.16)$$

Where \hat{z} , z are represented as predicted noise and true noise respectively,

$$\hat{z} = x - \hat{x}, \quad z = x - y \quad (5.17)$$

In this equation, α represents the ratio of energy between the clean signal (x) and the

noise (z), which can be written as

$$\alpha = \frac{\sum(x^2)}{\sum(x^2) + \sum(z^2)} \quad (5.18)$$

The above loss function (5.16) is used for the $N2C$ method. For $N2N$ and $N2N_{avg}$, in equation (5.16), x will be replaced with y_1 and y_3 respectively. The estimated spectrogram ($\hat{X}(t, f)$) is the product of noisy spectrogram $Y(t, f)$ and complex mask $\hat{M}(t, f)$.

$$\hat{X}(t, f) = Y(t, f) \cdot \hat{M}(t, f) \quad (5.19)$$

The complex mask $\hat{M}(t, f)$ can be defined in terms of magnitude and phase components as

$$\hat{M}(t, f) = M(t, f)^{mag} \cdot M(t, f)^{phase} \quad (5.20)$$

Where the magnitude and phase of the complex mask is defined as follows:

$$M(t, f)^{mag} = \tanh(f_p(Y(t, f))), M(t, f)^{phase} = \frac{f_p(Y(t, f))}{|f_p(Y(t, f))|} \quad (5.21)$$

Where the $f_p(Y(t, f))$ is output of the neural network. A detailed explanation of the mask polar coordinates is given in (Choi *et al.* 2019). Finally, the enhanced time-frequency spectrogram ($\hat{X}(t, f)$) was then transformed into its time-domain waveform using an Inverse STFT.

5.2.2 Methodology for Training and Testing

Three different training techniques were used to train the DCUnet-10 model: First, as shown in Fig. 5.2(a), the model was trained with noisy input, and the target was trained with clean speech. In the second technique, known as N2N, inputs and the targets were trained by two different noisy data sets as shown in Fig. 5.2(b). Finally, in the proposed method (N2N_{avg}) the DCUnet-10 was trained with noisy inputs and targets with average noisy data (as mentioned in the methodology) as shown in Fig. 5.2(c). The training and the evaluation data utilized for the three methods are listed in Table 5.2. All the training models were trained using an Nvidia Tesla K80 GPU with batch size of two (with 10 epochs). To evaluate the practical situation (field test), the network trained the input with $y_1(x_{tr1} + N_1)$ and the target with $y_3(x_{tr1} + N_{avg})$, assuming that no clean (x_{tr1}) information is available, then tested with a different

dataset ($x_{tst2} + N_1$) as shown in Table 5.2.

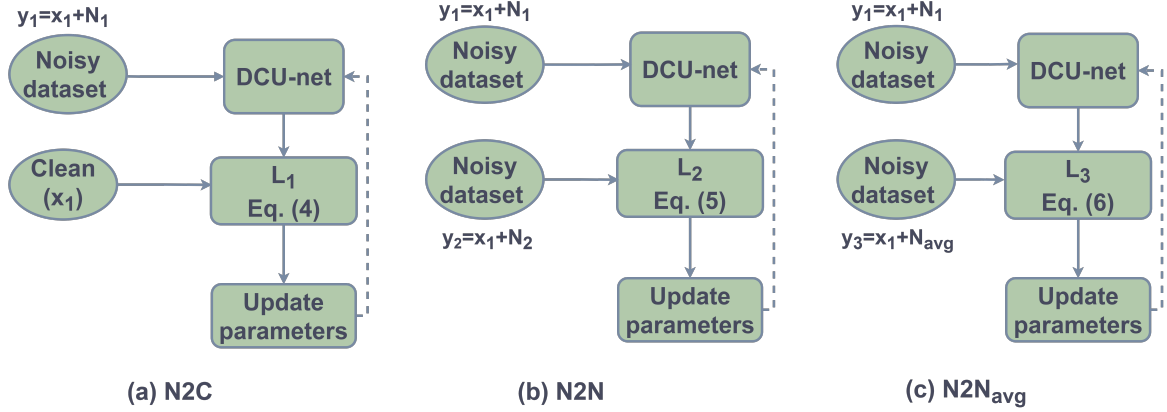


Figure 5.2: Overview of training methodologies.

Table 5.2: Methodology for training and evaluation

Method	Training		Testing	
	Input	Target	Matched	Mismatched
N2C	$x_{tr1} + N_1$	x_{tr1}	$x_{tst1} + N_1$	$x_{tst2} + N_1$
N2N	$x_{tr1} + N_1$	$x_{tr1} + N_2$	$x_{tst1} + N_1$	$x_{tst2} + N_1$
N2N _{avg}	$x_{tr1} + N_1$	$x_{tr1} + N_{avg}$	$x_{tst1} + N_1$	$x_{tst2} + N_1$

5.3 Experimental results and discussion

5.3.1 Dataset Generation

Table 5.3: Dataset generation

Dataset generation	Clean	Noise
Noisy Dataset I	Voice Bank dataset	+ N_1, N_2, N_{avg}
Noisy Dataset II	Voice Bank dataset	+ DEMAND
Noisy Dataset III	Open SLR66	+ UrbanSound8K

This work requires a noisy dataset to train the model both for input and the target. A noisy data set was generated with details shown in Table 5.3 as no standard data sets were available.

(a) Noisy Dataset I: For this study, a 28-speaker Voice Bank dataset (Valentini-Botinhao *et al.* 2016) was used which is a popular dataset used in most of the speech enhancement deep learning models (Koizumi *et al.* 2020 Kawanaka *et al.* 2020). In this dataset, 26 speakers (1452 samples) were used for training, and the remaining two (824 samples) were used for testing. This work used three noise categories, which are real-time English babble (N_1), and two synthetic noises (N_2, N_{avg}) for mixing with a clean dataset at 5 different SNR (-10, -5, 0, 5, 10) levels. N_{avg} was the average of N_1 and N_2 , where N_2 was stationary noise. Note that the noise classes were uniquely generated with two special conditions to train the deep neural network. The first condition was that all noise categories should be a zero mean distribution, and the second condition was that there should be no correlation between the two noises.

(b) Noisy Dataset II: Similarly, this study generated noisy data set by combining the Voicebank (Valentini-Botinhao *et al.* 2016) clean data with real-time noises from DEMAND (Thiemann *et al.* 2013) with different SNR levels (-5, 0, 5, 10, and 15 dB). The clean speech data is mixed with noise using the additive noise method, and the noisy dataset is generated using the MATLAB tool. This study compared the suggested strategy with N2C and N2N to determine whether N2N_{avg} enhances speech quality by using an average of noisy targets.

(c) Noisy Dataset III: This study tested the effectiveness of this training methodology on low-resource languages using the Open SLR66 data set (He *et al.* 2020). This dataset contains recordings of female and male Telugu native speakers. The female dataset contains 2294 sentences, of which 1720 were used for training and 574 for testing. The male dataset contains 2154 sentences, of which 1616 were used for training and 538 for testing. The noise sounds in the UrbanSound8K dataset (Salamon *et al.* 2014), which consists of several noise files lasting 4 sec each, were used to corrupt the clean speech signals. The synthetic noisy dataset was created by adding white Gaussian noise to clean speech samples, then using a taxonomy of urban sounds presented in (Salamon *et al.* 2014 & Wu *et al.* 2021). This work considered the four noise sounds most frequently encountered in urban environments: (i) Air Conditioning, (ii) Car horn, (iii) Children playing, and (iv) Engine idling. The clean speech was mixed with noise from Urbansound8K at different SNRs ranging from 0 to 10 dB. In all noisy speech samples, the sample rate is 48 kHz.

5.3.2 Objective parameters

To experimentally validate the effectiveness of the training methods, the estimated speech signal from the denoising network must be compared with the clean speech signal. The measurements used are introduced in the subsequent paragraphs and are dependent on clean speech which is assumed to be available only in the experimental setup and not in a real-world scenario.

5.3.2.1 Signal to distortion ratio (SDR) measures

SDR has been frequently employed as an objective measure and a classical measure of SNR. However, it requires both estimated (\hat{x}) and target (x) speech samples (Le Roux *et al.* 2019). This is how it might be calculated:

$$SNR = 10 \log_{10} \left(\frac{\sum_{k=1}^L (x_k)^2}{\sum_{k=1}^L (x_k - \hat{x}_k)^2} \right) \quad (5.22)$$

where x_k and \hat{x}_k are the clean and estimated samples of speech indexed by k and L represents the overall sample number.

Classical SNR and speech quality have a poor correlation, thus rather than taking into account the entire signal, SNR is determined as the mean of the SNRs of short segments (selected to be 30 ms) (Hu and Loizou 2007) and is defined as:

$$SegmentalSNR = \frac{10}{S} \sum_{s=0}^{S-1} 10 \log_{10} \left(\frac{\sum_{k=Ls}^{Ls+L-1} (x_k^2)}{\sum_{k=Ls}^{Ls+L-1} (x_k - \hat{x}_k)^2} \right) \quad (5.23)$$

where L is the number of samples in the segment and S is total number of segments.

5.3.2.2 Perceptual evaluation of speech quality (PESQ)

The PESQ (Recommendation 2001) algorithm predicts how people will rate the quality of degraded speech samples. It assigns a score between 4.5 and -0.5, where higher scores mean better quality.

5.3.2.3 Short time objective intelligibility (STOI)

The STOI metric is calculated by comparing the temporal envelopes of the time-aligned clean speech with its estimated speech signal in short-time overlapped seg-

ments (Taal *et al.* 2011).

5.3.2.4 Composite measures

Basic objective measurements including segmental SNR, weighted spectral slope (WSS), PESQ, and Log-likelihood ratio (LLR) are combined to create composite measures. PESQ is a metric used to determine subjective opinion scores for audio samples. Higher scores indicate greater quality, and PESQ delivers a value ranging from 4.5 to -0.5. Three composite measures (Hu and Loizou 2007) were used for this study: Composite measure for speech distortion (CSIG), Composite measure for overall signal quality (COVL), and Composite measure for noise distortion (CBAK). Higher values of these composite measurements, indicate that the method performs better. The results of their calculation with multiple linear regression analysis are as follows:

$$CBAK = 1.634 - 0.007 * WSS + 0.063 * SSNR + 0.478 * PESQ$$

$$CSIG = 0.603 * PESQ + 3.093 - 1.029 * LLR - 0.009 * WSS$$

$$COVL = 0.805 * PESQ + 1.594 - 0.007 * WSS - 0.512 * LLR$$

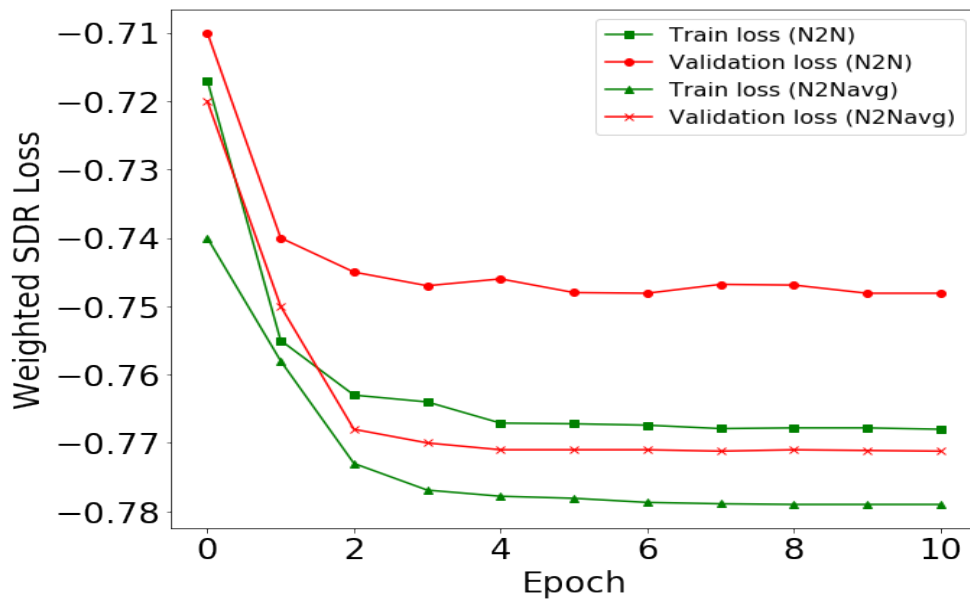


Figure 5.3: The training loss and validation loss.

Initially, to check whether or not the proposed method can provide a minimum loss

than the N2N, this study evaluated the loss ($Loss_{WSDR}$) of the training methods. The performance of the two methods in terms of training and validation losses is shown in Fig. 5.3. When compared to the N2N method, the proposed method (N2N_{avg}) had minimum training and validation losses. This study also observed that the difference between training and validation losses was smaller in N2N_{avg} than in N2N. Therefore, this shows that, compared to N2N, the N2N_{avg} method reaches N2C loss ($L1$) faster. Although the N2N_{avg} may require less data than N2N, it still needs more training data to achieve N2C. Hence, to avoid the high dependence on the massive data set in the proposed method, the average noise data set was given as the training output. This study hypothesized that the dependency of the data set size in the proposed method will be inversely proportional to the square of the number of noises added to form the average noise (as seen in section 5.2).

Table 5.4: Performance comparisons of different methods in terms of STOI and PESQ with the noisy dataset I

SNR (dB)	Method	PESQ	STOI
10	Noisy	2.023	0.812
	N2C	2.311	0.904
	N2N	2.494	0.853
	N2N _{avg}	2.5	0.878
5	Noisy	1.912	0.783
	N2C	2.031	0.803
	N2N	2.23	0.8009
	N2N _{avg}	2.25	0.8669
0	Noisy	1.623	0.713
	N2C	1.74	0.667
	N2N	1.969	0.729
	N2N _{avg}	2.00	0.76
-5	Noisy	1.01	0.521
	N2C	1.207	0.505
	N2N	1.621	0.54
	N2N _{avg}	1.632	0.672
-10	Noisy	0.67	0.443
	N2C	0.769	0.436
	N2N	1.237	0.575
	N2N _{avg}	1.239	0.618

The performance of the estimated speech was evaluated in terms of PESQ, STOI, and composite metrics for the following methods with noisy dataset I. Table 5.4 shows the intelligibility of the predicted speech by unprocessed (noisy), N2C, N2N, and N2N_{avg} (proposed method) with various SNR levels. In terms of PESQ and STOI scores, it was observed that the N2N_{avg} approach outperforms the N2C and N2N methods for the majority of SNR levels.

Table 5.5: Performance comparisons of N2C, N2N, and N2N_{avg} based on composite measures with the noisy dataset I

Method	CBAK	CSIG	COVL
Noisy	2.36	3.56	3.24
N2C	2.175	3.918	2.989
N2N	2.5606	3.76	3.437
N2N _{avg}	2.686	4.458	3.965

The effectiveness of the suggested method was determined using the following three composite measures: CSIG for signal distortion, CBAK for noise distortion, and COVL for overall quality as shown in Table 5.5. When it comes to composite measures, the N2N_{avg} approach performs better than N2N and N2C by providing higher scores. Higher scores indicate better reconstruction of the estimated signal.

The spectrogram of clean speech (Choi *et al.* 2019), the same speech corrupted by babble noise at 0 dB, improved speech with N2N, and speech fine-tuned by N2N_{avg} are shown in Fig. 5.4. As seen in Fig. 5.4(c), noisy training can effectively direct N2N in the elimination of background noise to some extent. On the other hand, Fig. 5.4(d) indicates that the N2N_{avg} can further reduce the residual noise with average noise training, increasing the PESQ score. This outcome confirmed the findings in Table 5.4 and Table 5.6 that N2N_{avg} can raise PESQ scores.

This work examined how the SNR of the noisy target affects N2N_{avg}'s performance since N2N_{avg} transforms into N2C when the SNR of the noisy signal (y) at the target equals that of the clean signal (x). To evaluate the effectiveness of the proposed method, this work employed the VoiceBank-DEMAND dataset. The SNR levels of the combined noisy targets were -5, 0, 5, 10, and 15 dB, and the noise signals were obtained from the DEMAND dataset. The objective evaluation was performed on VoiceBank-DEMAND to compare our method's performance with other cutting-edge approaches. This study used the VoiceBank-DEMAND (noisy dataset II) dataset for

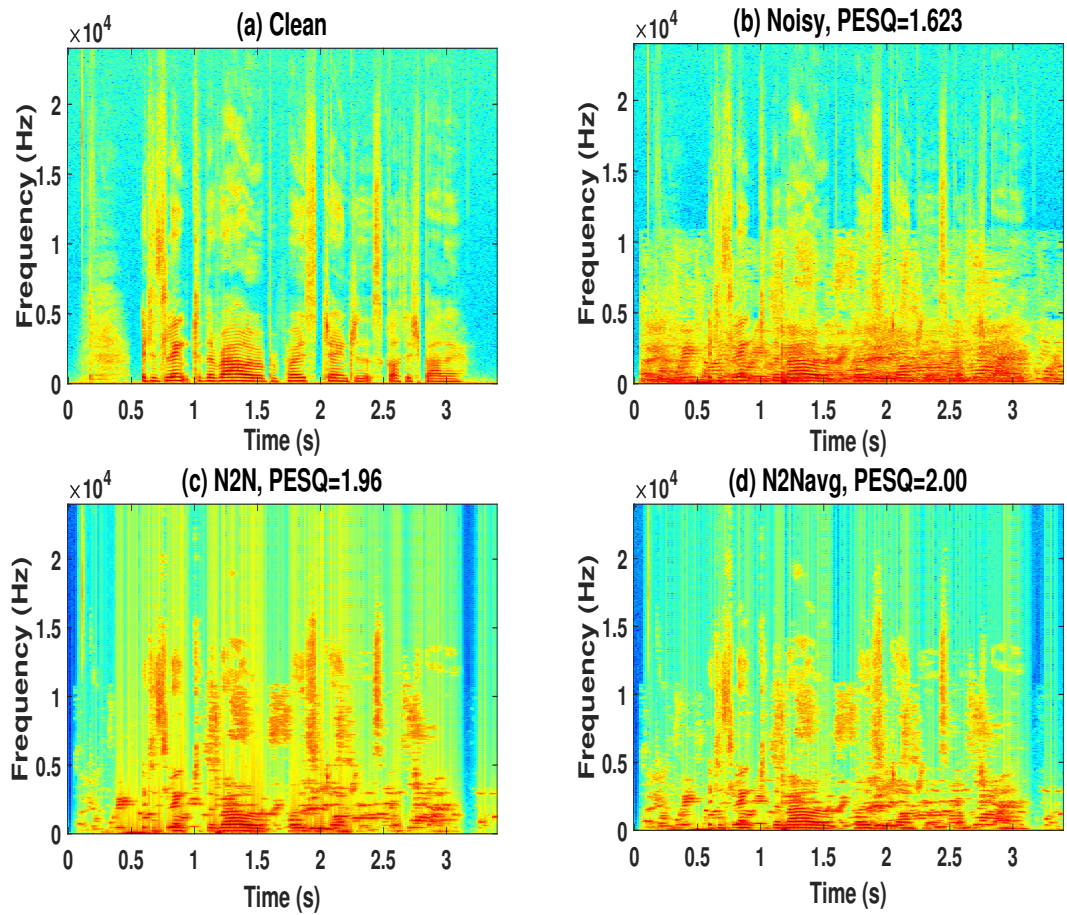


Figure 5.4: Spectrograms of a speech database (Valentini-Botinhao *et al.* 2016) sentence: (a) clean speech, (b) noisy speech (babble noise at 0 dB), (PESQ = 1.623), (c) denoised speech by N2N (PESQ = 1.96) (d) enhanced speech by N2N_{avg} (PESQ = 2.00).

evaluation to avoid the effect of the training/testing sample mismatch.

Table 5.6: Result of VoiceBank-Demand (without mismatch in training and testing data).

Method	Model	SDR	PESQ	STOI	CBAK	CSIG	CVOL
N2C	Noisy	8.97	1.912	0.67	2.36	3.50	2.98
	DCUnet (Choi <i>et al.</i> 2019)	18.88	2.43	0.73	3.18	3.75	3.24
	Segan (Pascual <i>et al.</i> 2017)	17.13	2.15	0.71	2.92	3.43	2.81
	Sergan (Baby and Verhulst 2019)	17.34	2.58	0.712	3.01	3.50	2.88
	Demucs (Defossez <i>et al.</i> 2020)	17.58	3.02	0.73	3.36	3.98	3.58
	CAUnet (Wang <i>et al.</i> 2021)	18.13	2.91	0.72	3.50	4.01	3.61
N2N	N ₁ 2N ₂ (Kashyap <i>et al.</i> 2021)	18.54	2.65	0.76	3.01	3.12	3.48
	N2N _{avg} (proposed)	18.56	2.71	0.78	3.11	3.78	3.57

Table 5.6 shows that in most metrics, the proposed method outperforms algorithms that train without a clean target (N₁2N₂) and few N2C algorithms. Even though N2C gives better results, our results have the advantage of the small dataset and don't require clean data set for training. In spite of that, results are close to N2C methods.

5.3.3 Validation of N2N_{avg} strategy

To evaluate the performance of the proposed training technique, this study conducted two experiments.

(a) signal-to-distortion ratio (SDR):

Signal-to-distortion ratio (SDR) (Hu and Loizou 2007) was first calculated to determine if the denoising network can be trained without clean speech. The evaluated SDRs are shown in Fig. 5.5, where N2N_{avg} receives better SNR than unprocessed speech (Noisy) with different noise conditions. The outcomes demonstrate that the speech-denoising network can be trained using the N2N_{avg} technique without clean speech.

(b) Subjective evaluations:

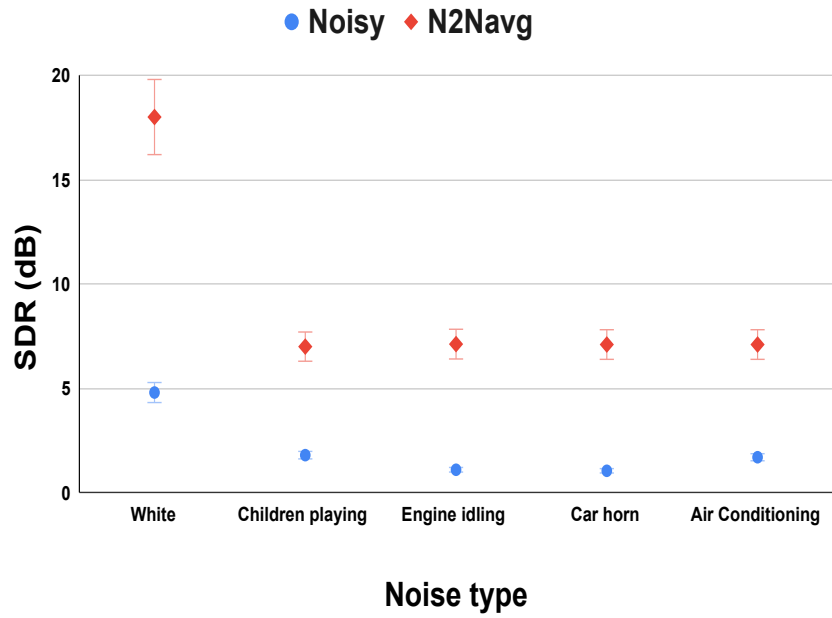


Figure 5.5: SDR comparison of noisy input signals and N2Navg results.

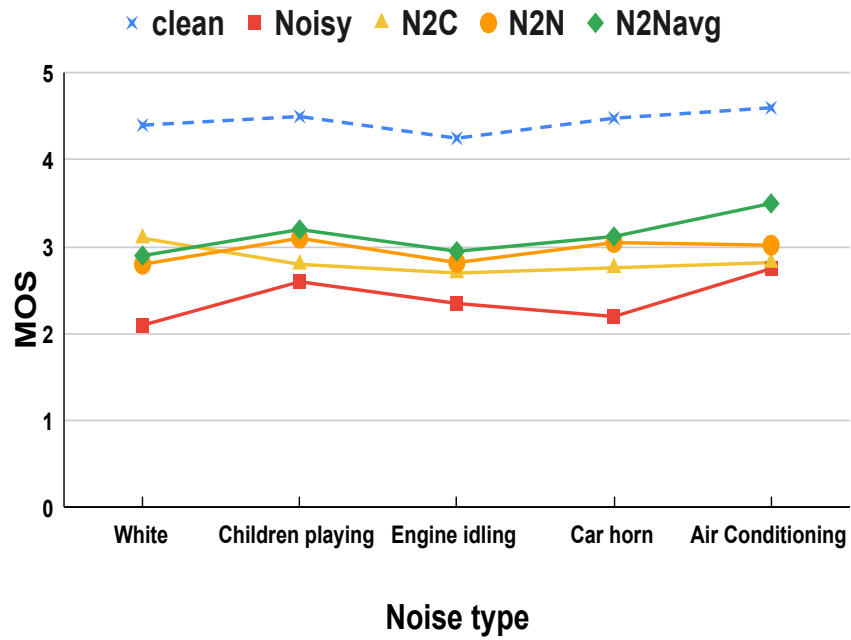


Figure 5.6: MOS evaluation results from a subjective perspective.

In the second experiment, to evaluate the speech intelligibility of the enhanced speech, a perceptual evaluation test was conducted. Twelve subjects with normal hearing were asked to rate the following five signals: (1) Clean, (2) Noisy (unprocessed), denoised using (3) N2C, (4) N2N, and (5) N2N_{avg}. Each subject heard a different combination of the above five audio signals for 15 speech sentences. Then, the individuals were asked to select their preferences for speech intelligibility preservation and noise suppression on a scale of 1 (very poor), 2 (poor), 3 (fair), 4 (good), and 5 (outstanding). Fig. 5.6 displays the subjective test results as an average among the 12 subjects. The outcomes of the subjective evaluation demonstrate that the proposed N2N_{avg} method achieves better mean opinion score (MOS) scores and exhibits the finest speech-denoising impact. For evaluating the performance of the N2N_{avg} with different noises, this study selected white noise, and four noises from UrbanSound8K, all five noises mixed with a TELUGU clean speech at different SNRs from 0 to 10 dB. The five noisy speech signals have been averaged to generate noisy target speech data. Based on the comparative experiments, Table 5.7 shows that the proposed method has superior performance compared to existing strategies (i.e., N2C, N2N) for all noises except White noise. It has also been shown that the proposed method offers comparative denoising performance compared to N2C in white noise, and also each metric exceeds two benchmark methods in all other noises.

Table 5.7: Evaluation of training methods with SLR TELUGU dataset

Noise type	Method	PESQ	STOI	SSNR
White	Noisy	1.526 ± 0.173	0.623 ± 0.003	3.51 ± 1.05
	N2C	2.66 ± 0.452	0.654 ± 0.004	3.88 ± 0.85
	N2N	2.497 ± 0.462	0.631 ± 0.005	2.96 ± 0.69
	N2N _{avg}	2.58 ± 0.467	0.67 ± 0.003	3.98 ± 0.61
Children playing	Noisy	1.42 ± 0.03	0.62 ± 0.03	-3.62 ± 1.04
	N2C	1.87 ± 0.14	0.64 ± 0.04	-1.08 ± 0.05
	N2N	1.98 ± 0.13	0.68 ± 0.05	-0.56 ± 0.51
	N2N _{avg}	2.12 ± 0.15	0.70 ± 0.003	0.17 ± 0.61
Car horn	Noisy	1.80 ± 0.250	0.79 ± 0.003	-3.82 ± 1.60
	N2C	2.15 ± 0.15	0.82 ± 0.003	1.78 ± 0.60
	N2N	2.18 ± 0.25	0.83 ± 0.003	1.98 ± 0.69
	N2N _{avg}	2.24 ± 0.459	0.85 ± 0.003	2.14 ± 0.68
Air Conditioning	Noisy	2.05 ± 0.18	0.841 ± 0.002	-2.68 ± 0.67
	N2C	2.46 ± 0.23	0.85 ± 0.003	2.58 ± 0.42
	N2N	2.47 ± 0.31	0.87 ± 0.003	2.96 ± 0.41
	N2N _{avg}	2.48 ± 0.32	0.88 ± 0.003	3.2 ± 0.45
Engine idling	Noisy	1.93 ± 0.45	0.78 ± 0.003	1.38 ± 0.48
	N2C	2.12 ± 0.23	0.75 ± 0.003	1.25 ± 0.38
	N2N	2.23 ± 0.38	0.81 ± 0.003	1.08 ± 0.41
	N2N _{avg}	2.31 ± 0.386	0.82 ± 0.003	2.06 ± 0.68

5.3.4 Evaluation of N2N_{avg} with mismatch condition

To prove the minimal dependence of the N2N_{avg} method on the clean dataset, this study has considered the mismatch condition also. To evaluate the performance of the proposed method, this study trained the model with the noisy dataset I ($x_{tr1} + N_1$ for input training, $x_{tr1} + N_{avg}$ for target training) and tested it with the SLR66 dataset (x_{tst2}) plus the noise (N_1) as shown in Table 5.2. Table 5.8 contains a summary of the mean scores. Although the performance is noticeably worse than matched conditions, the proposed method still outperforms N2C and N₁2N₂ in effectiveness. Hence, this is evidence that the model trained with noises but not on clean data.

Table 5.8: Result of SLR66 (x_{tst2})+ N_1 (with a mismatch in training and testing data).

Method	Training Model	SDR	PESQ	STOI	CBAK	CSIG	CVOL
Noisy	-	8.01	1.42	0.60	2.21	2.8	2.24
N2C	DCUnet(Choi <i>et al.</i> 2019)	12.18	1.73	0.64	2.55	3.28	2.57
N2N	N_12N_2 (Kashyap <i>et al.</i> 2021)	12.01	1.65	0.66	2.57	3.18	2.60
	$N2N_{avg}$ (proposed)	12.06	1.70	0.68	2.65	3.24	2.62

5.4 Summary

In this chapter, a novel self-supervised speech-denoising method is proposed. The method involves creating two distinct noisy datasets with different noise distributions that are uncorrelated with each other and have zero mean. By doing so, the proposed denoising method is able to effectively learn without requiring additional expensive clean data. The experimental results suggest that the proposed method outperforms other comparable strategies in most measurement metrics, indicating that it is a promising approach for speech denoising. Additionally, the proposed approach remains competitive even in situations where clean speech samples are limited, which is a common problem in speech denoising.

Overall, this study presents a significant contribution to the field of speech denoising by providing a new self-supervised method that overcomes the need for expensive clean data and outperforms existing techniques in many scenarios.

Chapter 6

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

The main goal of this thesis was to enhance speech clarity for people with cochlear implants. This was accomplished through two stages. Firstly, by encoding temporal fine structures (TFS) while considering the neuro-physiological constraints of cochlear implant users. Second, implementing noise reduction methods as pre-processing techniques to improve the signal-to-noise ratio for enhancing speech intelligibility in noise.

This thesis addressed various challenges in designing auditory filters and temporal fine structure (TFS) frequency compression within the neuro-psychological limitations of cochlear implant (CI) users. The application of proportional frequency compression to encode TFS within the temporal limits of CIs was investigated. Using a proportional frequency compression algorithm that encodes TFS, more harmonics were coded within the neuro-physiological imitations (300 Hz to 1000 Hz). Subjective evaluations were used to measure the speech recognition scores for all TFS coding conditions and the sinewave vocoder (No-TFS). The speech recognition threshold in noise (SRTN) was calculated from these speech recognition scores. The results showed that frequency-compressed TFS conditions had better SRTN than the sinewave vocoder. Notably, speech recognition scores with frequency-compressed 600 Hz TFS (300 Hz) surpassed those of the sinewave vocoder (NO-TFS), indicating a positive outcome. It is worth mentioning that the SRTN value was significantly lower (0.969 dB) when using frequency-compressed 600 Hz TFS (300 Hz) compared to the sinewave vocoder

(2.67 dB), which indicates a favourable result for temporal coding limited up to 300 Hz TFS in CIs. According to the Bayesian paired-sample T-test analysis, frequency-compressed TFS has proven beneficial. This suggests pitch shifting should be considered in future CI signal processing strategies. Neurophysiological studies focusing on pitch coding are recommended for prospective validation.

A modified Wiener filter method for noise reduction was proposed and compared with traditional Wiener filtering (WF) and sigmoidal functions using acoustic simulation of cochlear implants with normal hearing individuals. Compared to other methods, the proposed method offers the highest permissible gain for input signals having low SNR levels. The proposed method gives minimum MSE values when compared to WF. The SRTN was calculated for the proposed method, WF, and sigmoidal function based on the speech recognition scores of the volunteers. The required SRTN value for the proposed method was minimum (-3.167 dB) when compared to WF (0.167 dB) and sigmoidal function (0.23 dB). This indicated that the proposed method required minimum SNR for at least 50% speech perception compared to WF and sigmoidal function. After calculating the SRTN, statistical analysis was performed. A one-way ANOVA with repeated measurements ($p < 0.05$) was utilized to investigate the noise reduction effect. After analyzing the data, it was found that speech recognition with noise reduction (PM) is significantly better than without noise reduction (Unprocessed), with a p -value of 0.001. Additionally, the speech recognition with the WF was considerably better than the Unprocessed data, with a p -value of 0.039. This means that both PM and WF significantly improve speech recognition in the presence of speech-shape noise. However, the proposed method showed even more significant improvement in speech recognition results than the traditional Wiener filter, with a p -value of 0.035. The performance of the proposed method was evaluated using objective parameters such as ESTOI and SRMR-CI. Perceptual and objective analyses indicated that the proposed technique was more effective in improving speech intelligibility compared to sigmoidal functions and traditional Wiener filtering. This technique holds potential implications for CI applications, warranting further investigation with actual CI users.

Furthermore, a new self-supervised speech-denoising method was introduced, overcoming the need for expensive clean data and addressing the challenge of clean speech learning. By creating distinct noisy datasets with different, uncorrelated noise distributions, this work successfully tackled the need for clean data for training. The proposed method's performance was evaluated using both subjective and objective

evaluations. The proposed method demonstrated superior results with the Dataset I based on objective evaluations such as PESQ, STOI, and composite measures. The proposed method outperformed algorithms trained without a clean target (N2N) and a few N2C algorithms with the Voicebank-Demand (dataset II). Additionally, the proposed method's performance improved speech intelligibility, as quantified by PESQ, STOI, and SSNR metrics across various noise types (dataset III). Experimental results demonstrated the superiority of the proposed method over other comparative strategies across multiple metrics. Importantly, this approach remained competitive even when clean speech samples were limited, showcasing its potential compared to state-of-the-art techniques.

Overall, this thesis contributes to the advancement of auditory filter design, TFS frequency compression, noise reduction, and speech-denoising techniques for cochlear implants, opening avenues for further research and potential applications in the field.

6.2 Future Work

In this thesis, the study utilizes the acoustic simulation of the cochlear implant, not actual cochlear implants. The acoustic simulation revealed a positive benefit with proportional frequency compressed TFS and noise reduction methods. Hence, a similar method can be tested in an actual cochlear implant using electric stimulation. Also, the effect of training on speech recognition with frequency-compressed TFS is worthy of investigation.

Bibliography

- Alamdari, N., A. Azarang,** and **N. Kehtarnavaz** (2021). Improving deep speech denoising by noisy2noisy signal mapping. *Applied Acoustics*, **172**, 107631.
- Anderson, A. J.** and **A. J. Vingrys** (2001). Small samples: does size matter? *Investigative Ophthalmology & Visual Science*, **42**(7), 1411–1413.
- Apoux, F., C. L. Youngdahl, S. E. Yoho,** and **E. W. Healy** (2015). Dual-carrier processing to convey temporal fine structure cues: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, **138**(3), 1469–1480.
- Avinash, M., R. Meti,** and **U. Kumar** (2010). Development of sentences for quick speech-in-noise (quicksin) test in kannada. *J Indian Speech Hear Assoc*, **24**, 59–65.
- Azarang, A.** and **N. Kehtarnavaz** (2020). A review of multi-objective deep learning speech denoising methods. *Speech Communication*, **122**, 1–10.
- Baby, D.** and **S. Verhulst**, Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. *In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.
- Bianchi, F., L. H. Carney, T. Dau,** and **S. Santurette** (2019). Effects of musical training and hearing loss on fundamental frequency discrimination and temporal fine structure processing: Psychophysics and modeling. *Journal of the Association for Research in Otolaryngology*, **20**, 263–277.
- Chen, F., Y. Hu,** and **M. Yuan** (2015). Evaluation of noise reduction methods for sentence recognition by mandarin-speaking cochlear implant listeners. *Ear and hearing*, **36**(1), 61–71.

- Chen, J., J. Benesty, Y. Huang, and S. Doclo** (2006). New insights into the noise reduction wiener filter. *IEEE Transactions on audio, speech, and language processing*, **14**(4), 1218–1234.
- Chiea, R. A., M. H. Costa, and G. Barrault** (2019). New insights on the optimality of parameterized wiener filters for speech enhancement applications. *Speech Communication*, **109**, 46–54.
- Chiea, R. A., M. H. Costa, and J. A. Cordioli** (2021). An optimal envelope-based noise reduction method for cochlear implants: An upper bound performance investigation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 1729–1739.
- Choi, H.-S., J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee**, Phase-aware speech enhancement with deep complex u-net. *In International Conference on Learning Representations*. 2019.
- Chung, K.** (2007). Effective compression and noise reduction configurations for hearing protectors. *The Journal of the Acoustical Society of America*, **121**(2), 1090–1101.
- Crew, J. D. and J. J. Galvin III** (2012). Channel interaction limits melodic pitch perception in simulated cochlear implants. *The Journal of the Acoustical Society of America*, **132**(5), EL429–EL435.
- D’Alessandro, H. D., D. Ballantyne, P. J. Boyle, E. De Seta, M. DeVincenziis, and P. Mancini** (2018). Temporal fine structure processing, pitch, and speech perception in adult cochlear implant recipients. *Ear and hearing*, **39**(4), 679–686.
- De Boer, E. and H. De Jongh** (1978). On cochlear encoding: Potentialities and limitations of the reverse-correlation technique. *The Journal of the Acoustical Society of America*, **63**(1), 115–135.
- Defossez, A., G. Synnaeve, and Y. Adi** (2020). Real time speech enhancement in the waveform domain. *arXiv preprint arXiv:2006.12847*.
- Dhanasingh, A. and I. Hochmair** (2021). Signal processing & audio processors. *Acta Oto-Laryngologica*, **141**(sup1), 106–134.

- Ephraim, Y.** and **D. Malah** (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, **32**(6), 1109–1121.
- Ewert, S. D.** and **T. Dau** (2000). Characterizing frequency selectivity for envelope fluctuations. *The Journal of the Acoustical Society of America*, **108**(3), 1181–1196.
- Fischer, T., C. Schmid, M. Kompis, G. Mantokoudis, M. Caversaccio,** and **W. Wimmer** (2021). Effects of temporal fine structure preservation on spatial hearing in bilateral cochlear implant users. *The Journal of the Acoustical Society of America*, **150**(2), 673–686.
- Fu, S.-W., C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu,** and **Y. Tsao** (2021). Metricgan+: An improved version of metricgan for speech enhancement. *arXiv preprint arXiv:2104.03538*.
- Geetha, C., K. S. S. Kumar, P. Manjula,** and **M. Pavan** (2014). Development and standardisation of the sentence identification test in the kannada language. *J Hear Sci*, **4**(01), 18–26.
- Glasberg, B. R.** and **B. C. Moore** (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing research*, **47**(1-2), 103–138.
- Grais, E. M.** and **M. D. Plumbley**, Single channel audio source separation using convolutional denoising autoencoders. In *2017 IEEE global conference on signal and information processing (GlobalSIP)*. IEEE, 2017.
- Guo, N., S. Wang, R. Genov, L. Wang,** and **D. Ho** (2020). Asynchronous event-driven encoder with simultaneous temporal envelope and phase extraction for cochlear implants. *IEEE Transactions on Biomedical Circuits and Systems*, **14**(3), 620–630.
- Hast, A., L. Schlücker, F. Digeser, T. Liebscher,** and **U. Hoppe** (2015). Speech perception of elderly cochlear implant users under different noise conditions. *Otology & Neurotology*, **36**(10), 1638–1643.
- Hazrati, O., H. Ali, J. H. Hansen,** and **E. Tobey** (2015). Evaluation and analysis of whispered speech for cochlear implant users: Gender identification and intelligibility. *The Journal of the Acoustical Society of America*, **138**(1), 74–79.

- Hazrati, O., J. Lee, and P. C. Loizou** (2013). Blind binary masking for reverberation suppression in cochlear implants. *The Journal of the Acoustical Society of America*, **133**(3), 1607–1614.
- He, F., S. H. C. Chu, O. Kjartansson, C. E. Rivera, A. Katanova, A. Gutkin, I. Demirsahin, C. C. Johny, M. Jansche, S. Sarin, et al.** (2020). Open-source multi-speaker speech corpora for building gujarati, kannada, malayalam, marathi, tamil and telugu speech synthesis systems.
- Healy, E. W., S. E. Yoho, J. Chen, Y. Wang, and D. Wang** (2015). An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type. *The Journal of the Acoustical Society of America*, **138**(3), 1660–1669.
- Henry, F., M. Glavin, and E. Jones** (2021). Noise reduction in cochlear implant signal processing: A review and recent developments. *IEEE reviews in biomedical engineering*.
- Hersbach, A. A., K. Arora, S. J. Mauger, and P. W. Dawson** (2012). Combining directional microphone and single-channel noise reduction algorithms: a clinical evaluation in difficult listening conditions with cochlear implant users. *Ear and hearing*, **33**(4), e13–e23.
- Hjorungnes, A. and D. Gesbert** (2007). Complex-valued matrix differentiation: Techniques and key results. *IEEE Transactions on Signal Processing*, **55**(6), 2740–2746.
- Hochmair, I., P. Nopp, C. Jolly, M. Schmidt, H. Schöber, C. Garnham, and I. Anderson** (2006). Med-el cochlear implants: state of the art and a glimpse into the future. *Trends in amplification*, **10**(4), 201–219.
- Hu, Y. and P. C. Loizou** (2007). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, **16**(1), 229–238.
- Hu, Y. and P. C. Loizou** (2010). Environment-specific noise suppression for improved speech intelligibility by cochlear implant users. *The Journal of the Acoustical Society of America*, **127**(6), 3689–3695.

- Hu, Y., P. C. Loizou, N. Li, and K. Kasturi** (2007). Use of a sigmoidal-shaped function for noise attenuation in cochlear implants. *The Journal of the Acoustical Society of America*, **122**(4), EL128–EL134.
- Jensen, J. and C. H. Taal** (2016). An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(11), 2009–2022.
- Joris, P. X. and T. C. Yin** (1992). Responses to amplitude-modulated tones in the auditory nerve of the cat. *The Journal of the Acoustical Society of America*, **91**(1), 215–232.
- Kan, A. and Q. Meng** (2020). The temporal limits encoder as a sound coding strategy for bilateral cochlear implants. *IEEE/ACM transactions on audio, speech, and language processing*, **29**, 265–273.
- Kan, A., C. Stoelb, R. Y. Litovsky, and M. J. Goupell** (2013). Effect of mismatched place-of-stimulation on binaural fusion and lateralization in bilateral cochlear-implant users. *The Journal of the Acoustical Society of America*, **134**(4), 2923–2936.
- Kashyap, M. M., A. Tambwekar, K. Manohara, and S. Natarajan** (2021). Speech denoising without clean training data: A noise2noise approach. *arXiv preprint arXiv:2104.03838*.
- Kawanaka, M., Y. Koizumi, R. Miyazaki, and K. Yatabe**, Stable training of dnn for speech enhancement based on perceptually-motivated black-box cost function. *In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.
- Kim, G., Y. Lu, Y. Hu, and P. C. Loizou** (2009). An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, **126**(3), 1486–1494.
- Koizumi, Y., K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi**, Speech enhancement using self-adaptation and multi-head self-attention. *In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

- Koning, R., I. C. Bruce, S. Denys, and J. Wouters** (2018). Perceptual and model-based evaluation of ideal time-frequency noise reduction in hearing-impaired listeners. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **26**(3), 687–697.
- Koning, R., N. Madhu, and J. Wouters** (2014). Ideal time-frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners. *IEEE Transactions on Biomedical Engineering*, **62**(1), 331–341.
- Lai, Y.-H., F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee** (2016). A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation. *IEEE Transactions on Biomedical Engineering*, **64**(7), 1568–1578.
- Le Roux, J., S. Wisdom, H. Erdogan, and J. R. Hershey**, Sdr-half-baked or well done? *In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.
- Lehtinen, J., J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila** (2018). Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*.
- Li, X., K. Nie, L. Atlas, and J. Rubinstein**, Harmonic coherent demodulation for improving sound coding in cochlear implants. *In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.
- Li, X., K. Nie, N. S. Imennov, J. T. Rubinstein, and L. E. Atlas** (2013). Improved perception of music with a harmonic based algorithm for cochlear implants. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **21**(4), 684–694.
- Loizou, P. C.** (2006). Speech processing in vocoder-centric cochlear implants. *Cochlear and brainstem implants*, **64**, 109–143.
- Loizou, P. C.**, *Speech enhancement: theory and practice*. CRC press, 2013.

- Lorenzi, C., G. Gilbert, H. Carn, S. Garnier, and B. C. Moore** (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences*, **103**(49), 18866–18869.
- Lu, X., Y. Tsao, S. Matsuda, and C. Hori**, Speech enhancement based on deep denoising autoencoder. *In Interspeech*, volume 2013. 2013.
- Lu, Y. and P. C. Loizou** (2010). Estimators of the magnitude-squared spectrum and methods for incorporating snr uncertainty. *IEEE transactions on audio, speech, and language processing*, **19**(5), 1123–1137.
- Madhu, N., A. Spriet, S. Jansen, R. Koning, and J. Wouters** (2012). The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses. *IEEE Transactions on Audio, Speech, and Language Processing*, **21**(1), 63–72.
- Martin, R.** (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on speech and audio processing*, **9**(5), 504–512.
- Mauger, S. J., P. W. Dawson, and A. A. Hersbach** (2012). Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction. *The Journal of the Acoustical Society of America*, **131**(1), 327–336.
- McKay, C. M. and H. J. McDermott** (1996). The perception of temporal patterns for electrical stimulation presented at one or two intracochlear sites. *The Journal of the Acoustical Society of America*, **100**(2), 1081–1092.
- Meng, Q., N. Zheng, and X. Li**, A temporal limits encoder for cochlear implants. *In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.
- Meng, Q., N. Zheng, and X. Li** (2016). Mandarin speech-in-noise and tone recognition using vocoder simulations of the temporal limits encoder for cochlear implants. *The Journal of the Acoustical Society of America*, **139**(1), 301–310.

- Mesnildrey, Q., G. Hilkhuisen, and O. Macherey** (2016). Pulse-spreading harmonic complex as an alternative carrier for vocoder simulations of cochlear implants. *The Journal of the Acoustical Society of America*, **139**(2), 986–991.
- Micheyl, C. and A. J. Oxenham** (2010). Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. *Hearing research*, **266**(1-2), 36–51.
- Micheyl, C. and A. J. Oxenham** (2012). Comparing models of the combined-stimulation advantage for speech recognition. *The Journal of the Acoustical Society of America*, **131**(5), 3970–3980.
- Moon, I. J. and S. H. Hong** (2014). What is temporal fine structure and why is it important? *Korean journal of audiology*, **18**(1), 1.
- Moore, B. C.** (2021). Effects of hearing loss and age on the binaural processing of temporal envelope and temporal fine structure information. *Hearing research*, **402**, 107991.
- Moulines, E. and F. Charpentier** (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, **9**(5-6), 453–467.
- Mourao, G. L., M. H. Costa, and S. Paul** (2020). Speech intelligibility for cochlear implant users with the mmse noise-reduction time-frequency mask. *Biomedical Signal Processing and Control*, **60**, 101982.
- Müller, V., H. Klünter, D. Fürstenberg, H. Meister, M. Walger, and R. Lang-Roth** (2018). Examination of prosody and timbre perception in adults with cochlear implants comparing different fine structure coding strategies. *American Journal of Audiology*, **27**(2), 197–207.
- Müller, V., H. D. Klünter, D. Fürstenberg, M. Walger, and R. Lang-Roth** (2020). Comparison of the effects of two cochlear implant fine structure coding strategies on speech perception. *American Journal of Audiology*, **29**(2), 226–235.
- Nambi, P. M. A., Y. Mahajan, N. Francis, and J. S. Bhat** (2016). Temporal fine structure mediated recognition of speech in the presence of multitalker babble. *The Journal of the Acoustical Society of America*, **140**(4), EL296–EL301.

- Ngamkham, W., C. Sawigun, S. Hiseni, and W. A. Serdijn**, Analog complex gammatone filter for cochlear implant channels. *In Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. IEEE, 2010.
- Nie, K., L. Atlas, and J. Rubinstein**, Single sideband encoder for music coding in cochlear implants. *In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008.
- Nie, K., G. Stickney, and F.-G. Zeng** (2004). Encoding frequency modulation to improve cochlear implant performance in noise. *IEEE transactions on biomedical engineering*, **52**(1), 64–73.
- Nugraha, A. A., A. Liutkus, and E. Vincent** (2016). Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(9), 1652–1664.
- Pascual, S., A. Bonafonte, and J. Serra** (2017). Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*.
- Patterson, R. D., I. Nimmo-Smith, J. Holdsworth, and P. Rice**, An efficient auditory filterbank based on the gammatone function. *In a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, volume 2. 1987.
- Patterson, R. D., K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand**, Complex sounds and auditory images. *In Auditory physiology and perception*. Elsevier, 1992, 429–446.
- Peeters, G.** (1998). Analyse et synthèse des sons musicaux par la méthode psola. *Proceedings of the Journées d’Informatique Musicale (JIM)*.
- Plapous, C., C. Marro, L. Mauuary, and P. Scalart**, A two-step noise reduction technique. *In 2004 IEEE international conference on acoustics, speech, and signal processing*, volume 1. IEEE, 2004.
- Plapous, C., C. Marro, and P. Scalart** (2006). Improved signal-to-noise ratio estimation for speech enhancement. *IEEE transactions on audio, speech, and language processing*, **14**(6), 2098–2108.

- Poluboina, V., A. Pulikala, and A. N. P. Muthu** (2022). Contribution of frequency compressed temporal fine structure cues to the speech recognition in noise: An implication in cochlear implant signal processing. *Applied Acoustics*, **189**, 108616.
- Purdy, S. C., D. Welch, E. Giles, C. L. A. Morgan, R. Tenhagen, and A. Kuruvilla-Mathew** (2017). Impact of cognition and noise reduction on speech perception in adults with unilateral cochlear implants. *Cochlear implants international*, **18**(3), 162–170.
- Recommendation, I.-T.** (2001). Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*.
- Remus, J. J. and L. M. Collins** (2005). The effects of noise on speech recognition in cochlear implant subjects: predictions and analysis using acoustic models. *EURASIP Journal on Advances in Signal Processing*, **2005**, 1–12.
- Riss, D., J.-S. Hamzavi, M. Blineder, C. Honeder, I. Ehrenreich, A. Kaider, W.-D. Baumgartner, W. Gstoettner, and C. Arnoldner** (2014). Fs4, fs4-p, and fsp: A 4-month crossover study of 3 fine structure sound-coding strategies. *Ear and Hearing*, **35**(6), e272–e281.
- Ronneberger, O., P. Fischer, and T. Brox**, U-net: Convolutional networks for biomedical image segmentation. *In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015.
- Salamon, J., C. Jacoby, and J. P. Bello**, A dataset and taxonomy for urban sound research. *In Proceedings of the 22nd ACM international conference on Multimedia*. 2014.
- Santos, J. F., S. Cosentino, O. Hazrati, P. C. Loizou, and T. H. Falk** (2013). Objective speech intelligibility measurement for cochlear implant users in complex listening environments. *Speech communication*, **55**(7-8), 815–824.
- Santos, J. F. and T. H. Falk** (2014). Updating the srmr-ci metric for improved intelligibility prediction for cochlear implant users. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**(12), 2197–2206.

- Schnell, N., G. Peeters, S. Lemouton, P. Manoury, and X. Rodet**, Synthesizing a choir in real-time using pitch synchronous overlap add (psola). *In ICMC*. 2000.
- Serizel, R., M. Moonen, B. Van Dijk, and J. Wouters** (2014). Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22**(4), 785–799.
- Slaney, M. et al.** (1993). An efficient implementation of the patterson-holdsworth auditory filter bank. *Apple Computer, Perception Group, Tech. Rep*, **35**(8).
- Sohn, J., N. S. Kim, and W. Sung** (1999). A statistical model-based voice activity detection. *IEEE signal processing letters*, **6**(1), 1–3.
- Spahr, A. J., M. F. Dorman, and L. H. Loiselle** (2007). Performance of patients using different cochlear implant systems: effects of input dynamic range. *Ear and Hearing*, **28**(2), 260–275.
- Stronks, H., J. Briaire, and J. Frijns** (2020). The temporal fine structure of background noise determines the benefit of bimodal hearing for recognizing speech. *Journal of the Association for Research in Otolaryngology*, **21**(6), 527–544.
- Su, J., Z. Jin, and A. Finkelstein** (2020). Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. *arXiv preprint arXiv:2006.05694*.
- Taal, C. H., R. C. Hendriks, R. Heusdens, and J. Jensen** (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(7), 2125–2136.
- Takahashi, N., N. Goswami, and Y. Mitsufuji**, Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation. *In 2018 16th International workshop on acoustic signal enhancement (IWAENC)*. IEEE, 2018.
- Tejani, V. D. and C. J. Brown** (2020). Speech masking release in hybrid cochlear implant users: Roles of spectral and temporal cues in electric-acoustic hearing. *The Journal of the Acoustical Society of America*, **147**(5), 3667–3683.

- Teng, X., G. B. Cogan, and D. Poeppel** (2019). Speech fine structure contains critical temporal cues to support speech segmentation. *NeuroImage*, **202**, 116152.
- Thiemann, J., N. Ito, and E. Vincent**, The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. *In Proceedings of Meetings on Acoustics ICA2013*, volume 19. Acoustical Society of America, 2013.
- Trabelsi, C., O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal** (2017). Deep complex networks. *arXiv preprint arXiv:1705.09792*.
- Tseng, R.-Y., T.-W. Wang, S.-W. Fu, C.-Y. Lee, and Y. Tsao** (2020). A study of joint effect on denoising techniques and visual cues to improve speech intelligibility in cochlear implant simulation. *IEEE Transactions on Cognitive and Developmental Systems*, **13**(4), 984–994.
- Turner, C. W., B. J. Gantz, C. Vidal, A. Behrens, and B. A. Henry** (2004). Speech recognition in noise for cochlear implant listeners: benefits of residual acoustic hearing. *The Journal of the Acoustical Society of America*, **115**(4), 1729–1735.
- Valentini-Botinhao, C., X. Wang, S. Takaki, and J. Yamagishi**, Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks. *In Interspeech*, volume 8. 2016.
- Van Dijk, B., M. Moonen, J. Wouters, et al.** (2012). Speech understanding performance of cochlear implant subjects using time–frequency masking-based noise reduction. *IEEE transactions on biomedical engineering*, **59**(5), 1364–1373.
- Venema, T.** (1999). Three ways to fight noise: Directional mics, dsp algorithms, and expansion. *The Hearing Journal*, **52**(10), 58–60.
- Wang, D. and J. H. Hansen** (2018). Speech enhancement for cochlear implant recipients. *The Journal of the Acoustical Society of America*, **143**(4), 2244–2254.
- Wang, K., B. He, and W.-P. Zhu**, Caunet: Context-aware u-net for speech enhancement in time domain. *In 2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021.

- Wang, N. Y.-H., H.-L. S. Wang, T.-W. Wang, S.-W. Fu, X. Lu, H.-M. Wang, and Y. Tsao** (2020). Improving the intelligibility of speech for simulated electric and acoustic stimulation using fully convolutional neural networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **29**, 184–195.
- Williamson, D. S., Y. Wang, and D. Wang** (2015). Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, **24**(3), 483–492.
- Wouters, J., H. J. McDermott, and T. Francart** (2015). Sound coding in cochlear implants: From electric pulses to hearing. *IEEE Signal Processing Magazine*, **32**(2), 67–80.
- Wu, J., Q. Li, G. Yang, L. Senhadji, and H. Shu** (2021). Self-supervised speech denoising using only noisy audio signals. *arXiv preprint arXiv:2111.00242*.
- Xu, Y., J. Du, L.-R. Dai, and C.-H. Lee** (2013). An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters*, **21**(1), 65–68.
- Xu, Y., J. Du, L.-R. Dai, and C.-H. Lee** (2014). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **23**(1), 7–19.
- Zeng, F.-G.** (2002). Temporal pitch in electric hearing. *Hearing research*, **174**(1-2), 101–106.
- Zerem dini, J., M. A. B. Messaoud, and A. Bouzid** (2017). Multiple comb filters and autocorrelation of the multi-scale product for multi-pitch estimation. *Applied Acoustics*, **120**, 45–53.
- Zhao, H., S. Zarar, I. Tashev, and C.-H. Lee**, Convolutional-recurrent neural networks for speech enhancement. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
- Zhou, H., N. Wang, N. Zheng, G. Yu, and Q. Meng** (2020). A new approach for noise suppression in cochlear implants: A single-channel noise reduction algorithm. *Frontiers in Neuroscience*, **14**, 301.

Publications Based on the Thesis

Journals:

1. **Poluboina Venkateswarlu**, Aparna Pulikala, and Arivudai Nambi Pitchai Muthu. "Contribution of frequency compressed temporal fine structure cues to the speech recognition in noise: An implication in cochlear implant signal processing." *Applied Acoustics* 189 (2022): 108616.
2. **Poluboina Venkateswarlu**, Aparna Pulikala, and Arivudai Nambi Pitchai Muthu. "An Improved Noise Reduction Technique for Enhancing the Intelligibility of Sinewave Vocoded speech: Implication in Cochlear Implants." *IEEE Access* (2022).

Conferences:

1. **Venkateswarlu Poluboina**, Aparna Pulikala, and Arivudai Nambi Pitchaimuthu. "Cochlear Acoustic Model that Improves the Speech Perception in Noise by Encoding TFS." *Advances in VLSI, Communication, and Signal Processing: Select Proceedings of VCAS 2021*. Singapore: Springer Nature Singapore, 2022. 627-634.
2. Uma prasanna vishnu, **Venkateswarlu Poluboina**, Aparna Pulikala. "Speech Intelligibility Enhancement for Cochlear Implant using Multi Objective Deep Denoising Auto encoder." *INDICON-2023* (Accepted).

Journals Communicated:

1. **Poluboina Venkateswarlu**, Aparna Pulikala, and Arivudai Nambi Pitchai Muthu. "Speech enhancement with deep learning: Minimization of loss function in the absence of clean data" *Circuits, Systems & Signal Processing* (Under review).

Bio-data (FOR PHD)

Name : POLUBOINA VENKATESWARLU
Address : 2-192, Chilakalamarri, Anatasagaram, Nellore, Andhra pradesh 524302
Email : venkipoluboina92@gmail.com
Qualification : Mtech.