CODING STRATEGIES FOR COCHLEAR IMPLANTS UNDER ADVERSE ENVIRONMENTS

by

QUDSIA TAHMINA

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Engineering

at

The University of Wisconsin-Milwaukee

May 2016

ABSTRACT

CODING STRATEGIES FOR COCHLEAR IMPLANTS UNDER ADVERSE ENVIRONMENTS

by

QUDSIA TAHMINA

The University of Wisconsin-Milwaukee, 2016
Under the Supervision of Professor Yi Hu

Cochlear implants are electronic prosthetic devices that restores partial hearing in patients with severe to profound hearing loss. Although most coding strategies have significantly improved the perception of speech in quite listening conditions, there remains limitations on speech perception under adverse environments such as in background noise, reverberation and band-limited channels, and we propose strategies that improve the intelligibility of speech transmitted over the telephone networks, reverberated speech and speech in the presence of background noise. For telephone processed speech, we propose to examine the effects of adding low-frequency and high- frequency information to the band-limited telephone speech. Four listening conditions were designed to simulate the receiving frequency characteristics of telephone handsets. Results indicated improvement in cochlear implant and bimodal listening when telephone speech was augmented with high frequency information and therefore this study provides support for design of algorithms to extend the bandwidth towards higher frequencies. The results also indicated added benefit from hearing aids for bimodal listeners in all four types of listening conditions. Speech understanding in acoustically reverberant environments is always a difficult task for hearing impaired listeners. Reverberated sounds consists of direct sound, early reflections and late reflections. Late reflections are known to be detrimental to speech intelligibility. In this study, we propose a reverberation

suppression strategy based on spectral subtraction to suppress the reverberant energies from late reflections. Results from listening tests for two reverberant conditions (RT_{60} = 0.3s and 1.0s) indicated significant improvement when stimuli was processed with SS strategy. The proposed strategy operates with little to no prior information on the signal and the room characteristics and therefore, can potentially be implemented in real-time CI speech processors. For speech in background noise, we propose a mechanism underlying the contribution of harmonics to the benefit of electroacoustic stimulations in cochlear implants. The proposed strategy is based on harmonic modeling and uses synthesis driven approach to synthesize the harmonics in voiced segments of speech. Based on objective measures, results indicated improvement in speech quality. This study warrants further work into development of algorithms to regenerate harmonics of voiced segments in the presence of noise.

To
my parents,
my husband,
and especially my sons

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACE - Advanced Combination Encoder

BP - Bandpass filter

CA - Compressed Analog Strategy

CI - Cochlear Implant

CIS - Continuous Interleaved Sampling Strategy

CNC - Consonant-Nucleus-Consonant

EAS - Electric and Acoustic Stimulation

F0 - Fundamental Frequency

HA - Hearing Aids

HP - Highpass Filter

HR - Harmonic regeneration

LP - Lowpass Filter

NR - Noise reduction

PSTN - Public Switched Telephone Networks

SAS - Simultaneous Analog Strategy

SPEAK -Spectral peak

WB - Wideband

ACKNOWLEDGMENTS

I would like to express my appreciation and gratitude to my advisor, Dr.Yi Hu for his suggestions and continuous support during this dissertation work. I am very thankful for his encouragement and all the research opportunities he provided me in the field of cochlear implants. I am glad to have worked in the Cochlear Implant and Auditory prosthesis lab and proud to be the first doctorate graduating from this lab. While working on this dissertation, I learned to be determined and persistent in my efforts, which enabled me to explore strategies and techniques to improve the performance of cochlear implants. I believe this research work has provided guidance and support for the state of the art cochlear implant coding strategies for technological advancements.

I am grateful to my committee members, Dr. Chiu Tai law, Dr. Jun Zhang, Dr. Seyed Hossieni and Dr. Yin Wang for their precious time and their valuable comments and suggestions in thesis proposal defense. All their suggestions had contributed to improve my final dissertation. Dr. Chiu Tai Law has served as the Department Chair of Electrical Engineering and has been very helpful and supportive throughout my studies in many different ways. I had the opportunity to assist him with one of his courses on Analytical Methods in Engineering, which provided me with a deeper understanding on time and frequency domain analysis and transform techniques. I took Digital Signal Processing course taught by Dr. Zhang, which was very helpful for my dissertation in clarifying the concepts of sampling, temporal and spectral analysis and filter design techniques. I learned a huge deal of speech signal processing from this course.

Dr. Seyed Hossieni has served as a Department Chair of Computer Science and has expressed his interest in my research and always supported me. I truly thank Dr Yin Wang for

serving on the dissertation committee and providing useful advice and feedback on this dissertation.

I should also acknowledge that my dissertation was supported in part by NIDCD/NIH (Grant No. R03 – DC008887) and Research Growth Initiative (RGI) grant from the University of Wisconsin-Milwaukee. I had the pleasure of working with my lab mates, Behnam Azimi, and Moulesh Bhandary. They were always helpful and supportive.

I am very grateful to my parents for their encouragement, which has always motivated to achieve my goals in every step of my life. All that I am today, is because of my parent tireless efforts in nurturing me to obtain higher education and dreaming a better life for me. Even though both of my parents worked, they were able to dedicate their time and efforts to educate us. Without their support, I would not have come this far in pursuing my dream. I would like to thank my brothers and sister in motivating me in every step of my life and supporting me.

Finally, I am grateful for having a caring and helpful husband Mr. Mohammad Khan, who has firmly and continuously supported and encouraged me. He had always been there for me and sacrificed a lot to make this dream come true. Now that I am finishing my PhD studies, he will soon be able to concentrate on his professional career. My children are a sense of comfort and cheer of my life. They have enabled me to be strong and confident in my abilities. I hope this work would set an example for them to achieve their dreams and be successful in their career and lives. Thank you all for supporting me throughout this journey.

Milwaukee, WI

December 14, 2015

CHAPTER 1

INTRODUCTION

Speech is the most natural mode of communication for human beings. Perception of speech has a functional significance on daily activities in human life. The auditory system employs cognitive, motor and sensory processes to hear and understand speech. Profound sensorineural hearing loss can severely impact personal, work and social life. Cochlear Implants (CI) are auditory prosthetic devices that provide the opportunity for people with profound hearing impairment to recover partial hearing. Cochlear implants serves as a useful means to convey auditory sensation by means of an array of electrodes that are surgically implanted into the cochlea of the inner ear. These electrodes bypasses the damaged parts of the auditory system and electrically stimulates the auditory-nerve fibers that sends the signals to the brain. Cochlear implants provide representations of the frequency information that are not adequately amplified by the hearing aids. Originally, cochlear implant devices were designed as single channel [9]; however the technological developments in engineering and improvements in speech processor designs has led to the evolution of multi-channel implants. These multi-channel implants have an advantage of electrically stimulating several nerve fibers in the cochlea using various frequency bands, thereby transmitting more detailed information to the brain. Due to the technological advancements and ongoing research studies in the area of cochlear implants, several coding strategies have been developed with a primary goal to improve the perception of speech.

1.1 Background

Cochlear implants allow completely deaf individuals to perceive the sensation of sound and allow them to perform regular activities. Benefits of implantation have been observed in recipients' subjective descriptions [21] and distinction between phonemes, words and sentences [51, 80, 104].

Many research studies primarily focused on improving the perception of speech with cochlear implants in quiet environments. As a result of these studies, several coding strategies were developed over the years that present the importance of preserving the temporal envelope, waveform and spectral information for cochlear implant listening.

Speech communication in day-to-day life occurs under a range of different environmental conditions. Although speech recognition in quiet listening conditions show a clear auditory benefit, presenting materials in adverse environments is often challenging for CI listeners. In transit from speaker to listener, speech signals are often modified by interfering signals such as background noise, reverberation, and by imperfections of the frequency or temporal response of the communication channel [3]. One major factor in these situations appears to be the lack of spectral and temporal fine structure information and resolution in cochlear implant signal processing strategies. Three different adverse environments have been identified, where cochlear implant users have difficulty understanding speech: telephone communication network, acoustic reverberation and presence of background noise.

1.2 Motivation

Normal hearing listeners can easily understand conversation over telephone. However, perception of telephone speech is difficult for many hearing impaired listeners, including cochlear implant listeners, due to the reduced frequency range, lack of visual cues and reduced audibility of telephone signals. Extending the bandwidth of telephone speech could potentially improve the performance of cochlear implants. Most existing studies [26, 47, 61, 101] focused on the performance comparison of broadband and band-limited speech perception but never differentiated the band-limiting effects derived from low frequency and high frequency. In this dissertation, we access and differentiate the contributions from added low- and high- frequency

information to the perception of telephone processed speech in hearing impaired listeners to provide guidance for the development of bandwidth extension techniques.

Acoustic reverberation is a phenomenon of enclosed spaces that can negatively impact the performance of cochlear implants. A speech signal captured by a human ear or a microphone placed at a distance from the source is smeared by reverberation due to the reflections from the walls, floors, ceilings or furniture. Reverberation flattens the formant transitions in vowels, fills the gaps and silent intervals associated with vocal tract closure in stop consonants, blurs the onset and offset of syllables, smear spectral cues and reduces temporal amplitude modulations [3, 6, 57, 76, 77] which results in reduced speech recognition and speech intelligibility scores. Several dereverberation algorithms were proposed to cope with the adverse impacts of reverberation [34, 55, 56, 61, 73]. Most of these algorithms focused on retrieving the direct sound from reverberated speech. In any enclosed space, reverberant sounds heard by the listener are comprised of direct sound, early reflections and late reflections. Early reflections are considered to be beneficial to intelligibility whereas, late reflections are detrimental. Due to the different perceptual effects of early and late reflections on human auditory system, there is a need to differentiate between the two types of signals. In this dissertation, we propose a strategy to suppress the reverberant energies from late reflections to improve the performance of cochlear implants under acoustic reverberant conditions. In this method, we propose to suppress the reverberant energies caused by late reflections from reverberated speech.

Speech communication always takes place in the presence of some form of background noise. Environmental sounds such as traffic noise, conversations in the background, mechanical noise from appliances and devices such as air conditioning, refrigerator and computer are common forms of interferences. Background noise reduces the signal-to-ratio (SNR) as portions of the

signal are made inaudible or corrupted by interfering signals. Although, normal hearing listeners can understand speech in the presence of white noise, each types of interference in the form of noise masker has different effects on speech intelligibility due to many different factors [20, 84]. Understanding speech in presence of background noise have always been difficult task for cochlear implant listeners [3, 9, 20, 24, 25, 42, 83, 110]. The increased difficulty in understanding speech in noise is due to the reduced audibility of speech, weakly conveyed F0 cues due to the poor spectral resolution and lack of temporal fine structure [18, 42]. Numerous studies have been conducted on noise reduction methods to enhance the perception of noisy speech such as spectral subtraction, short-time spectral amplitude estimation, wiener filtering, adaptive noise canceling and subspace method [6, 19, 39, 41, 67]. The goal of the noise reduction algorithm is to improve the quality and intelligibility of noisy speech. We propose a method that relies on harmonic regeneration after noise reduction to further enhance the intelligibility of speech. In this method, we explore the potential benefits of suppressing noise and then regenerating the harmonics in the voiced frames. We will be evaluating the contribution of harmonics by synthesizing the voiced segments of the speech. This approach is designed to make the beneficial cues from the harmonic model to be more salient to the cochlear implant listeners.

1.3 Proposed Work

In this dissertation, we propose coding strategies to enhance the perception of speech by cochlear implant listeners in three adverse environments: telephone communication network, reverberation and background noise.

We aim to:

- Develop coding strategies to enhance the speech intelligibility and perception of cochlear implant listeners in adverse environments.

- Develop a strategy to assess the effects of adding low- and high- frequency information to the band-limited telephone speech. We will also investigate whether the hearing aids benefits the recognition of telephone speech by cochlear implant listeners, which could be implemented by studying the combined electric and acoustic stimulation in bimodal listeners.
- Develop and efficient reverberation suppression strategy to suppress the additive reverberant energies from late reflections. The goal is to use spectral subtraction strategy to reduce overlap masking effects caused by late reflections.
- Apply harmonic regeneration technique after noise reduction to enhance the perception of noisy speech. The aim is to establish mechanism underlying the contributions of harmonics to the benefit of electric and acoustic stimulation. We propose a method based on harmonic modeling that uses synthesis driven approach to synthesize the harmonics in voiced segments of speech.

This dissertation is organized as follows. Chapter 2 provides the overview of the cochlear implant technology and coding strategies implemented for CI speech processors. The benefits of acoustic hearing from hearing aids, electric hearing from cochlear implants, and the interesting concept of combined electric and acoustic stimulations are also presented in Chapter 2. Chapter 3 gives a comprehensive review of the literature of the strategies developed in the area of telephone processed speech, acoustically reverberated speech and speech in the presence of background noise. Chapter 4 present the proposed strategy for assessing the effects of adding low and high frequency information for telephone processed speech. It also presents the evaluation of added benefits from hearing aids to the CI perception of telephone speech by studying the combined electric and acoustic stimulations. In Chapter 5, the proposed

reverberation suppression strategy based on spectral subtraction (SS) method is presented. This chapter also presents the comparison of spectral subtraction strategy with ideal reverberant masking strategy in terms of CI subjective performance. Chapter 6 presents noise reduction technique based on harmonic regeneration. A synthesis driven approach based on harmonic regeneration after noise reduction is explained and the objective measures for speech quality are also discussed in this chapter. Chapter 7 concludes with a summary of our contributions and future work in the field of cochlear implants.

CHAPTER 2

COCHLEAR IMPLANTS AND SIGNAL PROCESSING

Hearing loss is a common health problem, with more than three hundred and sixty million people worldwide having disabling hearing loss. The principal cause of the sensorineural hearing loss is a damage to the inner ear (cochlea) or to the sensorineural hair cells (that provides a pathway to the brain). This is the most common type of permanent hearing loss. Loss of sensorineural hearing may result from illness, aging, exposure to loud noise, use of certain toxic drugs, and also genetic causes. The Cochlear Implant is considered to be the most successful of all the neural prosthesis in terms of restoration of function. Cochlear implants are medical prosthetic devices that consists of electrode array, which is surgically implanted into the cochlea and electrically stimulate the surviving auditory neurons. Cochlear implants can help people with moderate to profound hearing loss in one or both ears and those who receive little or no benefit from hearing aids. These electronic devices have provided hearing ability to more than three hundred thousand individuals worldwide. It was shown that people with cochlear implant achieve an average of 80% sentence understanding, compared with 10% sentence understanding for hearing aids [2]. For several years, researchers, doctors and scientists have been attempting to restore hearing in profoundly deafened individuals. Cochlear implant technology has been rapidly developing for the last two decades as a result of which multichannel cochlear implants have been introduced to replace the single channel devices to provide improved intelligibility and recognition of sounds. Research and development in engineering and technology have resulted in advanced criteria for cochlear implant candidacy to include congenitally deaf children for multichannel cochlear implantation. Due to the technological developments, multichannel cochlear implants can now be used in prelingually

deafened children as young as 12 months of age. It can be seen that cochlear implant technologies, as well as the outcome of the implantation are continuously evolving, thus enhancing the hearing ability of hearing impaired.

2.1 Human Auditory System

Hearing is the ability to perceive sound by detecting vibrations through the ear. The human auditory system consists of sensory organs and auditory peripherals which are responsible for detecting and processing sensory information to provide a sense of hearing to an individual. The human ear has exquisite intensity and frequency resolution capabilities. The frequency range of the auditory system is between 20 Hz to 20 kHz with ordinary conversation range of 300 Hz to 3 kHz and dynamic range of human hearing is about 120 dB. The human auditory periphery is divided into four sensory units: outer ear, middle ear, inner ear and auditory nerves as shown in the Figure 2.1. Sound waves travel through the outer ear and get modulate by the middle ear and excites the vestibulocochlear nerves of the inner ear. The external portion is the outer ear, which consists of the auricle (pinna) and the ear canal. The tympanic membrane (ear drum) and three bones or ossicles (malleus, incus and stapes) forms the middle ear. The three bones are arranged so that movement of one causes the other to move and eventually results in movement of fluid within the cochlea. The ossicles help in amplification of sound waves and reduce the amount of sound reflection. The inner ear is the internal portion of the ear that contains sensory organ for hearing called the cochlea. The cochlea is a spiral shaped structure with three fluid-filled spaces scala tympani, scala vestibuli and scala media which are responsible for hearing sensation. The sound waves propagate from the base through the spiral canal (30mm long) of the cochlea to the apex. The sound waves coming from the middle ear cause the extracellular fluid in the cochlea to move in the basilar membrane. The motion of the fluid in the membrane is picked up by the hair

cells with the help of stereocilia which transforms the mechanical energy of sound waves into electrical signals which ultimately leads to an excitation of the auditory nerves. These primary auditory neurons transform the signals into electrochemical impulses known as action potential, which travel along the auditory nerve to structures in the brainstem for further processing.

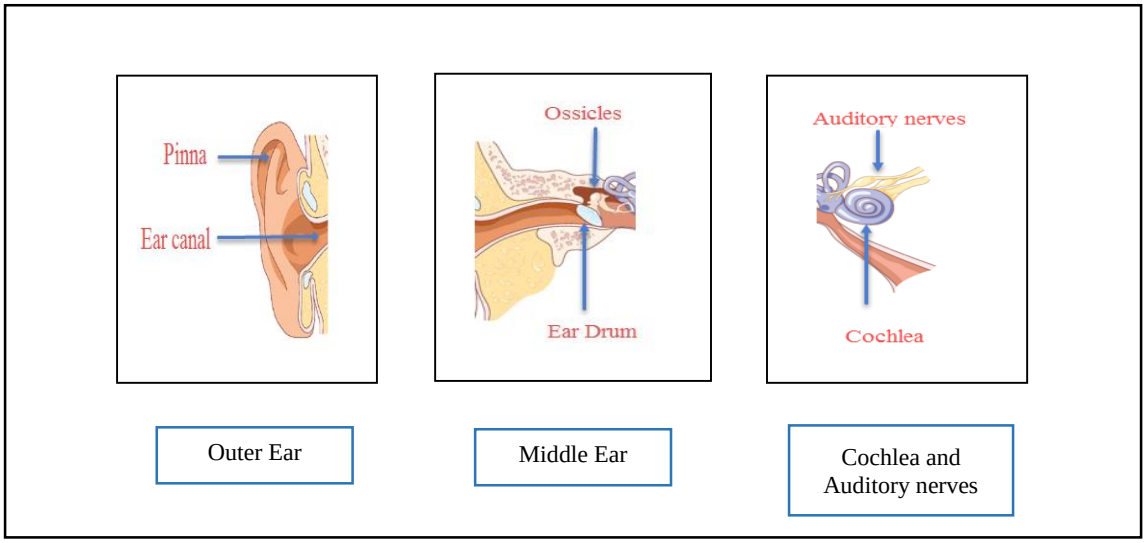


Figure 2.1. A Block diagram of human auditory system.

The basilar membrane acts as a filter and transmits parts of a sound waves into the auditory-nerve fibers due to which the cochlea appears to be a bank of filters that transmit information in parallel. These filters are called auditory filters with overlapping frequency ranges that perform spectral analysis.

2.2 Hearing loss and the need for cochlear implants

In normal hearing listeners, sound travels from the outer ear through the auditory canal and excites the tympanic membrane causing vibrations in the middle and inner ear. The pressure variations in the oval window cause the cochlear fluid to move in the basilar membrane. In response to the sound waves, the basilar membrane vibrates. Due to the tonotopic organization of the basilar membrane, each location along the basilar membrane responds best to one frequency. Although it

responds to other frequencies as well, this characteristic accounts for the spectral resolution of the human ear. The vibrations of the basilar membrane are amplified and compressed by the outer hair cells which provide level and frequency dependent gain control and aid in the exquisite sensitivity and frequency resolving capabilities of the ear. These vibrations cause the hair cells to bend in order to create potential difference that generates electrical currents. Finally, the transduction of the inner hair cells triggers the auditory nerve fibers that carry information to the brain.

Unfortunately, the hair cells are sensitive and fragile in nature, and can be damaged easily causing lack of transduction of the inner hair cells. The inner hair cells in the cochlea helps transmit the information to the auditory nerve fibers that sends the signals to the brain. When a large amount of inner hair cells are damaged, the person is said to be profoundly deaf or hearing impaired. In these cases, a hearing aid will not be helpful as it only amplifies the sound but there is no medium to transmit those signals to the brain due to the loss of inner hair cells. Therefore, an alternative was to use electrical stimulations to activate the inner hair cells that could send the signals to the brain. This implementation is introduced as cochlear implants, which are surgically implanted into the cochlea to regain the hearing sensations through electrical stimulation.

2.3 Brief history of cochlear implants

After intense research for several decades, an interest in biological application of electricity was the basis for the development of cochlear implants. The idea of using electrical stimulation to activate the auditory system in individuals with profound sensorineural hearing loss dates back to the year 1800, when the inserted metal rods in Alessandro Volta's ear canal created auditory sensations. In 1957, Djourno and Eyries reported a clear auditory percept by placing a wire carrying electrical current on the auditory nerve an individual undergoing surgery. Similar results were reported in 1961, when House and Doyle tested two profoundly deaf individuals by

electrically stimulating their auditory nerves using an electrode placed into the scala tympani of the inner ear [9]. As the level and rate of delivered stimulation changed, the loudness and pitch of stimuli also changed. Later in another study, by Simmons in 1964, identification of tone and changes in the sound duration were observed when an electrode was placed onto the modiolus of the cochlea. All these observations provided the basis for the development of functional and efficient cochlear implant systems. For those people who may receive little to no benefit from hearing aids, cochlear implantation has proven to be a viable option [3, 115].

2.4 Structure of cochlear implants

A cochlear implant device is made up of two parts, external portion that placed behind the ear and an internal portion that is surgically implanted under the skin. A cochlear implant device consists of the following parts:

- A microphone, which receives the surrounding sounds from the environment.
- A speech processor, which filters, analyzes and arranges the sounds received by the microphone into frequency bands.
- A transmitter, which transmits signals received from speech processor to the receiver placed underneath the skin.
- A receiver/stimulator, which receives the signals from the transmitter and sends the electrical impulses to the array of electrodes.
- An electrode array, which is a set of electrodes that activates/stimulates different regions of the auditory nerves around the cochlea.

An implant is not capable of restoring normal hearing, instead it can provide a representation of sounds received from the environment to help hearing impaired or deaf individual to understand speech. In case of single-channel implants, only one electrode is used, whereas for multi-channel

implants there are multiple electrodes that are used to stimulate the auditory nerves. The structure of the cochlear implant is shown in the Figure 2.2. The external microphone is used to pick the surrounding sounds and send them to the speech processor. The main function of speech processor is to filter the incoming signals into different frequency bands or channels similar to the way normal cochlea processes acoustic signals and delivering the filtered signals to the assigned electrodes. The signals from these frequency bands are sent to the transmitter and gets transmitted to the internal receiver using RF signals.

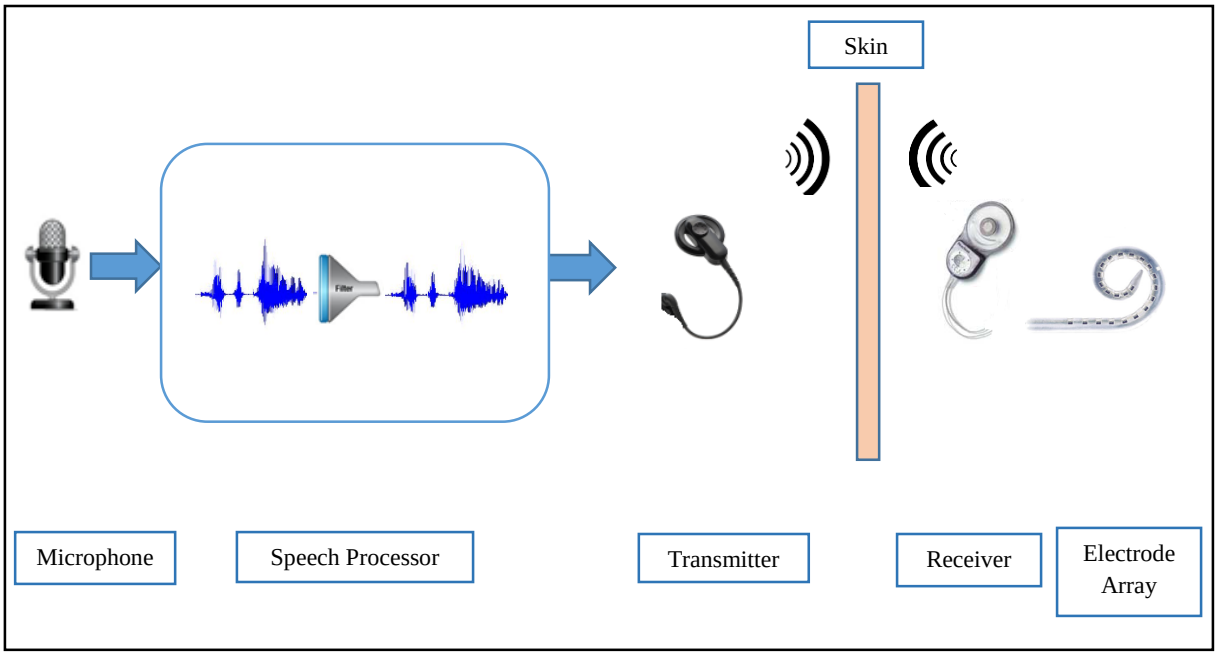


Figure 2.2. Structure of a Cochlear Implant device.

The receiver then converts them into electrical impulses and sends them to the electrode array which stimulates different regions of cochlea. These stimulations activate the auditory nerves that sends the signals to the brain and thus sounds are perceived by the brain.

2.5 Classification of cochlear implant devices

Cochlear implants have steadily gained popularity in the deaf community since their inception in early 1970's, and since then there were many technological and research developments in this field. Based on the electrodes used for stimulation, cochlear implant can be classified into two types: single-channel device or multi-channel device. If only one electrode is used for stimulation, then it is called a single-channel device. And if there are several electrodes used for stimulation, then it is a multi-channel device. Most of the early cochlear implants were single channel, but as the technology developed, multiple electrode were being used to increase the available frequency spectrum to the hearing impaired. These days almost all cochlear implants devices are multichannel devices. Usually, these are anywhere from 16 to 22 channels of stimulation depending on the individual's needs.

Other important aspect of classification are the stimulation type and transmission link. The electrical stimulation that drive the electrodes could be analog or pulsatile in nature. If the transmission link between the signal processor and the electrode array is a direct electrical connection, it is called a percutaneous link, and if it is based on radio frequency then it is called a transcutaneous link.

2.5.1 Single-channel cochlear implants

The first cochlear implant was implanted in human subjects in early 1970s, which was developed by William House and therefore named as 3M /House cochlear implant [23, 66]. The single channel implant consisted of an amplifier, a band-pass filter and a modulator. The incoming acoustic signals were first amplified and filtered using a single band-pass filter with the frequency range from 340-2700 Hz. The filtered signal was then modulated using a carrier frequency of 16 kHz. The modulated signal is then applied as input to an output amplifier whose gain can be varied

by the cochlear implant user. Although this single channel device provided many users with significant speech reading enhancement, it has limitations on the receiving end. The receiver does not perform demodulation and directly presents the high frequency signal to the single electrode as the stimulus. The single electrode could only simulate only a particular place in the cochlea which represents only limited frequency information for spectral analysis. Later in 1980's, another single channel cochlear implant named Vienna/3M was developed at the Technical University of Vienna, Austria [65].

2.5.2 Multi-channel cochlear implants

Multi-channel implants use an array of electrodes to provide electrical stimulations at multiple regions in the cochlea [65]. In 1980s, researchers were experimenting with variable number of electrodes to design an efficient implant system. A multichannel implant with 22 electrodes (Nucleus 22) was introduced by Cochlear Corporation which consisted of an implanted receiver/stimulator and an intra-cochlear electrode array to stimulate the nerve fibers. The multi-channel cochlear implants stimulated multiple auditory nerve fibers due to a large number of electrodes implanted at different locations along the cochlea. In these implants, radio frequency pulses were originally used to provide power for the implanted receiver and to control stimulation. Multi-channel electrodes present improved spectral analysis in a way similar to that performed by the auditory system via set of band-pass filters that exploit the frequency/place mechanism and provide better frequency resolution.

Another type of multichannel implant was developed with only 6 electrodes connected through a permanent percutaneous connector to the external speech processor called an Ineraid speech processor [9]. The processor utilizes a microphone, four bandpass filters and an analog electronic circuitry to provide output to each electrode. The percutaneous connector was used to

provide continuous analog signals to the intra-cochlear electrodes. The bandpass filtering and use of multiple electrode contacts was based on the theory that the brain would be able to extract the features of speech. It was shown from clinical trials with postlingually deafened adults that multichannel cochlear implant performed better than the single channel devices [27, 36].

Several researchers studied way to improve the design of implant system and identify the best intra-cochlear array size and stimulation mode. The idea was to reduce the size of the hardware for external and internal parts and also to fine tune the processing strategies. Depending on the type of information to be transmitted to the electrode (waveform derived filtering or spectral features such as first and second formants), researchers developed different types of signal processing techniques.

2.6 Commerical cochlear implant processors

Criteria for cochlear implant candidacy depends on the subject/participant criteria provided by the medical status of the patient. The guidelines are defined by the FDA based on clinical investigations on the safety and efficacy of the implants. There are three FDA approved manufacturers of multi-channel cochlear implant processors available in the United States [9]. The Nucleus processors are marketed by Cochlear Corporation and uses spectral peak (SPEAK) strategy. The Clarion devices are manufactured by Advanced Bionics Corporation and some of the processors used compressed analog (CA) strategy and some others used continuous interleaved sampling (CIS) strategy. The other processors were marketed by Medical Electronics Corporation (Med-El), which used high-rate continuous interleaved sampling (CIS) or a high-rate spectral peak (SPEAK) strategies. Over the past decade, all these processors has helped significantly improve the performance of the cochlear implant devices and helped hearing impaired listeners hear better.

The signal processing strategies used in the above mentioned processors are discussed in the next section.

2.7 Cochlear implant signal processing strategies

Signal processing strategies for multi-channel cochlear implants are basically divided into two categories based on how information is extracted from the speech signal and delivered to the electrodes. These two categories are: waveform based and feature-extraction based strategies [49, 65, 66]. The first is the waveform based approach, in which the signal is band-pass filtered and the corresponding filtered waveform is used to derive electric stimuli for the different electrodes. The second approach is based on feature extraction, in which important speech features like fundamental frequency and formant information are extracted using different algorithms and presented to the electrodes.

There are several factors that govern the presentation of acoustic stimuli to the electrodes that are determined by the signal processing strategies being used. Some of these factors are number of electrodes used for stimulation, electrode configuration, electrical current amplitude and compression functions. Other factors involved in the signal processing specific to the pulsatile stimulation are pulse rate and pulse width. Current cochlear implants use as many as 16-24 electrodes for stimulation. The frequency resolution of the cochlear implant device is determined by the number of electrodes used for stimulation and the neuron survival rate for each individual recipient of cochlear implant.

Various electrode configurations are used to control the power and currents provided to the electrodes since the current spread symmetrically into the electrode. With analog stimulations, there are two main kinds of electrode configurations used in the cochlear implant devices. These are mono-polar and bipolar configurations. In the mono-polar electrode configuration, a single

common ground is used for all the electrodes, which causes overlapping of the electric fields from neighboring electrodes and results in channel interaction. In the bipolar configuration, each individual electrode has its own ground without any overlapping of the electric fields and channel interaction. Other factors are compression of acoustical signal amplitude into electric current amplitude. In order to reduce the wide dynamic range of the acoustic sounds from the environment (30 – 50 dB) into narrow dynamic range of electrical impulses (about 5dB), a nonlinear mapping function is used. There are two ways to perform the compression of the acoustics signals: logarithmic function and a power-law function to obtain the electric current amplitudes. The larger the electric current amplitude, the louder the perceived stimulation. The smaller the amplitude of electric current the softer the perceived stimulation. In pulsatile stimulations, pulse rate determines the rate of stimulation of electrodes and pulse width determines the duration of single stimulation. The pulse shape can be monophasic or biphasic but mostly biphasic pulses are used for modulation.

2.7.1 Waveform based strategies

2.7.1.1 Compressed Analog and Simultaneous Analog Strategies

The compressed analog (CA) was developed by the researchers at Symbion, Inc., the company which manufactured the Ineraid cochlear implant. It is a waveform based strategy in which an automatic gain control (AGC) is computed to compress the signal before filtering. Four band-pass filters with frequency bandwidths of 100-700 Hz, 700-1400 Hz, 1400-2300 Hz, and 2300-5000 Hz with center frequencies of 500 Hz, 1000 Hz, 2000 Hz and 3400 Hz are applied to create four frequency bands/channels respectively [66]. The automatic gain control is then applied to the filtered signals and sent to the four intracochlear electrodes for stimulation. The compressed analog strategy provided better speech understanding to many patients and it was reported that the mean word identification was 14% for mono syllabic words and 45% for CID sentences [16]. But due to

the analog nature of the stimuli, it is continuously fed to all the electrodes at the same time, which results in channel interaction. Therefore, an extension to CA strategy was developed called Simultaneous Analog Strategy

Simultaneous Analog Strategy (SAS) provided continuous and simultaneous stimulation to the electrodes. Instead of four channels, this strategy used seven band-pass channels to provide more spectral information. After the automatic gain control, the signals are passed through a pre-emphasis and analog to digital converter to enhance the high frequency content. The signals are then band-pass filtered in digital domain and then multiplied by a gain factor similar to CA strategy. These signals are then compressed to fit into the electrical dynamic range using a controlled gain function. The stimuli is then presented to the electrodes in the analog form. The advantage of this strategy as that compression was customized for each user to optimize the performance.

2.7.1.3 Continuous Interleaved Sampling Strategy

Continuous Interleaved Sampling (CIS) strategy was developed by the researchers at Researchers at the Research Triangle Institute (RTI), which was also used to address the channel interaction issues with Compressed Analog (CA) strategy using non-simultaneous, interleaved pulses. In this strategy, biphasic pulse train is delivered to the various electrodes in a non-overlapping way [66]. One electrode is activated at a time, and the stimulation is cycled through various electrodes in a continuous way. The input signals from the microphone are pre-emphasized and filtered using bandpass filters. Using rectification and low-pass filtering the envelopes of the filtered waveforms are then extracted using full-wave rectification and low-pass filtering and are then the resulting channel envelopes are compressed using a non-linear compression mapping to fit the dynamic range of the electrical signals. These compressed signals are used to modulate biphasic pulses

which are sent to activate the electrodes as seen in Figure 2.3. Sequence of balanced biphasic pulses are delivered to the electrodes at a constant rate for both the voiced and unvoiced portion of the speech signal. Table 2.1 shows 3dB passband frequencies for 16 channel electrodes as an example. CIS strategy is used in the Clarion processors developed by the Advanced Bionics Corporation and Nucleus CI24M device.

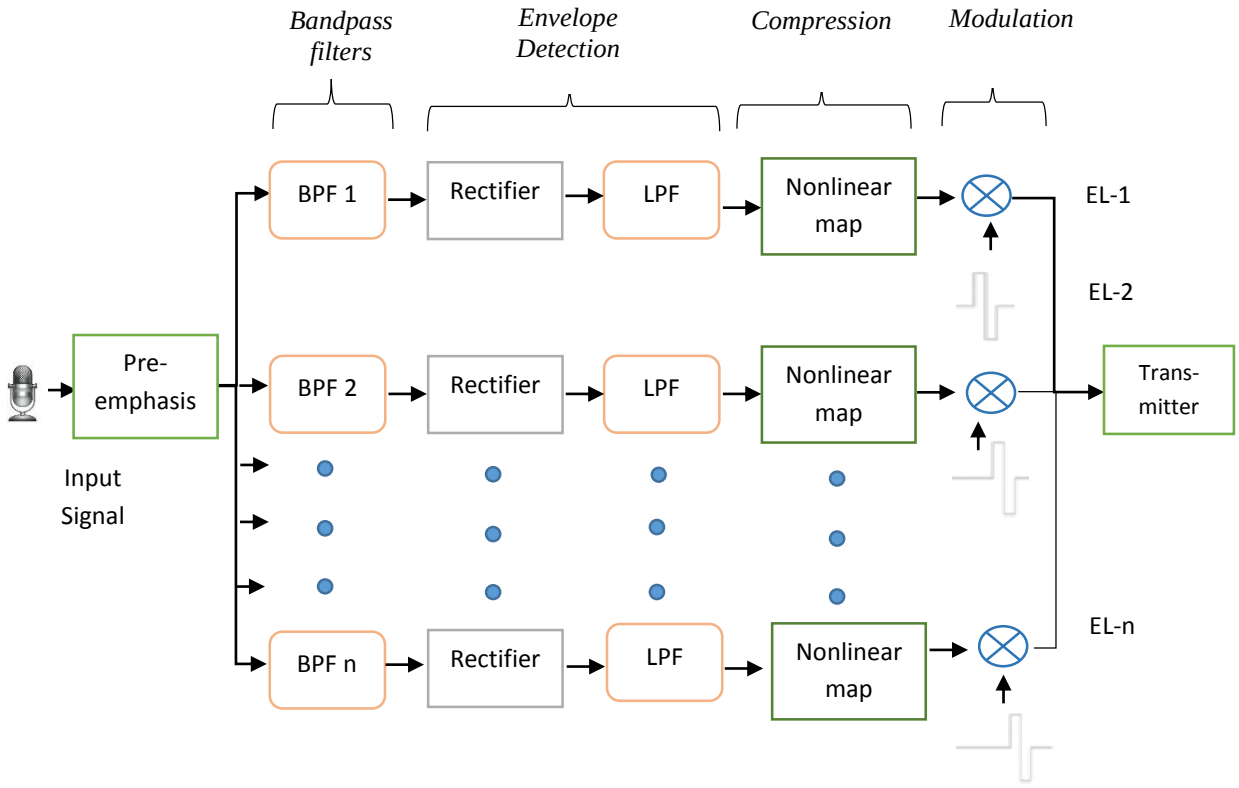


Figure 2.3. Signal Processing in Cochlear Implant devices (CIS Strategy).

2.7.1.4 SPEAK Strategy

Compared to other strategies, Spectral peak (SPEAK) strategy uses a 20-channel band-pass filter bank to perform the spectral analysis. The input signals are first filtered using bandpass filters and the channel amplitudes are detected using amplitude detection module. For each channel amplitude, spectral maxima is obtained by comparing it with the base value.

	Lower Band	Center	Upper Band
Channel 1	350 Hz	382.87 Hz	415.75 Hz
Channel 2	415.75 Hz	454.80 Hz	493.86 Hz
Channel 3	493.86 Hz	540.25 Hz	586.64 Hz
Channel 4	586.64 Hz	641.74 Hz	696.85 Hz
Channel 5	696.85 Hz	762.31 Hz	827.77 Hz
Channel 6	827.77 Hz	905.52 Hz	983.28 Hz
Channel 7	983.28 Hz	1075.64 Hz	1168.01 Hz
Channel 8	1168.01 Hz	1277.72 Hz	1387.44 Hz
Channel 9	1387.44 Hz	1517.77 Hz	1648.10 Hz
Channel 10	1648.10 Hz	1802.91 Hz	1957.72 Hz
Channel 11	1957.72 Hz	2141.62 Hz	2325.52 Hz
Channel 12	2325.52 Hz	2543.96 Hz	2762.41 Hz
Channel 13	2762.41 Hz	3021.90 Hz	3281.38 Hz
Channel 14	3281.38 Hz	3589.62 Hz	3897.85 Hz
Channel 15	3897.85 Hz	4263.99 Hz	4630.14 Hz
Channel 16	4630.14 Hz	5065.07 Hz	5500 Hz

Table 2.1. 3-dB pass-band frequencies for 16 channel electrode used for CIS strategy.

The channel amplitudes greater than the base value are used to stimulate the corresponding electrodes in a tonotopic order. Only electrodes corresponding to the spectral maxima are stimulated from base to apex. The pulse rate of the stimuli varies due to different number of electrodes stimulated in each cycle. Nucleus Spectra 22 processor utilizes this signal processing strategy.

2.7.1.5 ACE Strategy

Another strategy similar to SPEAK strategy was developed that utilized Fast Fourier Transform (FFT) to perform filtering of the input signal as described [49]. Input signals are filtered using a 128 point FFT at a sampling frequency of 16 kHz to produce frequency channels. The envelopes are extracted for each frequency channel using a low-pass filter with cut-off frequency of 180 Hz. The pulse rate in each frequency channel is adjusted to deliver the stimuli at a constant rate. The best feature of this strategy was that the stimuli could be delivered in two ways: SPEAK approach

(stimuli delivered to selected approach) or CIS approach (stimuli delivered to all electrodes). Nucleus 24 processors utilized the ACE signal processing strategy.

2.7.2 Feature extraction strategies

2.7.2.1 F0/F2 Strategy

The F0/F2 strategy was the first feature extraction strategy developed in 1980s. Speech features such as fundamental frequency (F0) and formant information are considered important for speech recognition due to the fact that they convey information and prosody of speech sounds. In this strategy, a low-pass filter is used with a cutoff frequency of 270 Hz along with a zero crossing detectors to extract the fundamental frequency (F0). And then a bandpass filter is used with frequencies between 1000 Hz and 4000 Hz along with another zero-crossing detector to extract the second formant (F2) information from the speech signal. By stimulating the selected electrode at F0 pulses/sec, voicing information was conveyed in voiced segments, whereas for unvoiced segments, information was conveyed by stimulating electrodes with an average rate of 100 pulses/sec at quasi-random intervals [66]. F0/F2 strategy was used in the Nucleus processors.

2.7.2.2 F0/F1/F2 Strategy

F0/F2 was modified to include the first formant information and therefore F0/F1/F2 strategy was developed. In addition to two zero crossing detectors used in F0/F2 strategy, another zero crossing detector was used to extract F1 information from the bandpass filtered (280 Hz to 1000 Hz) signal [66]. F0/F1/F2 strategy uses F0 to convey voicing information and two pulses in each time cycle to convey information about first and second formants to two corresponding implanted electrodes respectively. Out of the 20 electrode channels, the first five apical electrodes were used for transmitting information of the first formant information and the remaining fifteen electrodes were

used for transmitting second formant information. Therefore, two electrodes are stimulated at a time, one corresponding to the first formant (F1) and the other corresponding to the second formant (F2) with the pulse rate coding the fundamental frequency (F0). This strategy was implemented in the Nucleus wearable speech processor.

2.7.2.3 MPEAK Strategy

Improvements to F0/F1/F2 strategy in terms of new hardware and refinement of signal processing led to the development of MPEAK strategy. In order to improve the consonant recognition and enhance the representation of second formant, this strategy used up to three additional band-pass filters to provide high frequency information in addition to fundamental frequency (F0), first formant (F1) and second formant (F2) information. The envelope amplitudes are estimated for frequency bands 2000-2800 Hz, 2800-4000 Hz and 4000-6000 Hz [66]. Stimulation occurs on the electrodes corresponding to first two formants (F1 and F2) and on the high-frequency electrodes corresponding to 2000-2800 Hz and 2800-4000 Hz for voiced segments. Since there is not enough energy in the voiced spectrum above 4000 Hz, the electrode with 4000 – 6000 Hz does not get stimulated. And, for unvoiced segments, stimulation occurs on the electrode corresponding to second formant (F2) as well as the three high-frequency electrodes: 2000-2800 Hz, 2800-4000 Hz and 4000-6000 Hz. Since there is little energy in unvoiced sounds below 1000 Hz, the electrode corresponding to F1 does not get stimulated.

2.8 Electric and Acoustic Stimulation

Cochlear implants are used to electrically stimulate the auditory nerves and replace the inner sensory hair cells that are lost. In case of profound hearing loss, few surviving hair cells that are present in the cochlea cannot perceive speech information through acoustic input, and therefore cochlear implants tend to effective treatment, where array of electrodes replace the functionality

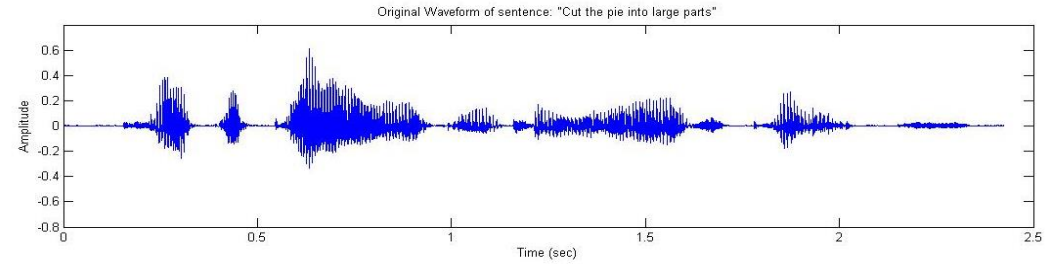
of inner hair cells across the frequency range. Cochlear implants present the middle and high frequencies, whereas the hearing aids acoustically amplify the low frequency information. Using hearing aids and cochlear implant simultaneously results in combined electric and acoustic stimulations (EAS). The concept of enhancing the hearing capabilities from combined electric and acoustic stimulations was first introduced by Von Ilberg *et al*. [105]. Individuals with low-frequency residual hearing in one or both ears can integrate electrically elicited percepts (E) from the implant and acoustically elicited percepts (A) from a hearing aid for speech understanding [59]. For patients with low frequency acoustic hearing, frequencies between 500 Hz to 750 Hz are available while higher frequencies of from 2000 Hz to 8000 Hz are provided by the electrical stimulations. Speech understanding is better with combined electric and acoustic stimulation (EAS) than with either E-alone or A-alone stimulation.

2.8.1 Hearing Aid benefit

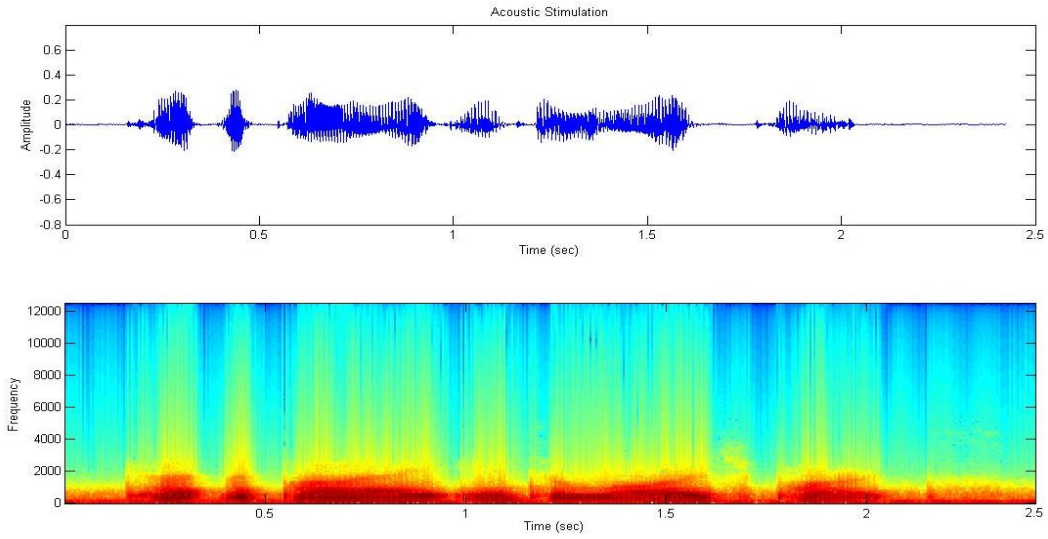
As more people are receiving cochlear implant, there has been more interest in exploring options for better speech recognition. Although cochlear implants provide speech understanding, there is a limited ability of cochlear implants to provide adequate frequency resolution. Researchers have identified that recipient of cochlear implants have various degrees of usable hearing in the contralateral ear. These patients have the opportunity of combining acoustic plus electric hearing. Residual acoustic hearing in the contralateral ear has potential benefits in speech understanding. Generally, patients implanted with a standard-length electrode in one ear, use a hearing aid in the contralateral ear.

Benefits of providing amplified speech to the contralateral ear have been reported by many researchers. It was also observed that compared with the cochlear implant alone listening, the combined condition achieved better speech recognition in backgrounds [59]. As it can be seen in

Figure 2.4, that the low frequency acoustic frequencies are made available through hearing aids. It was presumed that acoustic hearing assists in pitch perception and therefore capable of separating target voices from background. In another study, it was found that there were improvement in speech recognition from acoustic contralateral hearing to the implanted ear [29].



(a)



(b)

Figure 2.4. Effect of Acoustic stimulation on a speech signal. (a) Original speech spectrum of a sentence from IEEE database, (b) Acoustically stimulated speech spectrum.

It was also reported that speech recognition was comparable to the results obtained when the acoustic hearing was preserved in the implanted ear and listened in the A+E mode. In cases where the ear contralateral to the implant does not suffer any hearing loss, then improved speech recognition performance might not be linked to the combined acoustic plus electric hearing, and instead be primarily due to the listening from the better ear . Even after considering these factors, there is a strong evidence that addition of acoustic hearing can provide a substantial advantage for many hearing impaired listeners.

2.8.2 Cochlear implant benefits

In the case of profound or severe hearing loss, usually few functioning inner hair cells are present in the cochlea and very little usable speech information can be received through acoustic input to the cochlea thus hearing aids is not an effective treatment strategy. Electrical hearing from cochlear implants assist with the mid to high frequency information (see Figure 2.5) and that is why most of the cochlear implant listeners have difficulty with low frequency information.

For people with hearing loss, cochlear implants provide several benefits. Adults often seem to benefit immediately after the implantation compared to children. A significant amount of training and rehabilitation is required after implantation to help with the process of new electric hearing experienced by the cochlear implant users. Cochlear implant can help users make telephone calls and identify familiar voices over the telephone, although some good performers can understand unfamiliar voices as well. It improves the tone identifications and discrimination sound localization abilities of individuals and help them identify target speaker in the complex listening environments. There is significant improvement in music perception, identifying musical instruments and speech prosody. However, listening to the radio might pose challenges due to the lack of availability of visual cues.

Examples of improved performance include word, phoneme and sentence recognition [21, 51, 80, 104]. Variety of research studies have demonstrated improved scores on intelligibility in quiet, while perception of speech in noise and other factors has continued to be a challenge for implant users. Originally, cochlear implant surgery has been performed in one ear referred as a unilateral implantation. Because of the wide range of benefits from cochlear implants and enhanced life of individuals, researchers began to explore the benefits of bilateral implantation with speech understanding and localization. There are various demographic factors that might contribute to the results obtained with cochlear implants. Etiology, age at the time of implantation, age at the onset of deafness, duration of deafness, duration of implantation, neuron survival rate, prior auditory experience are some of the factors affecting the performance of cochlear implant users.

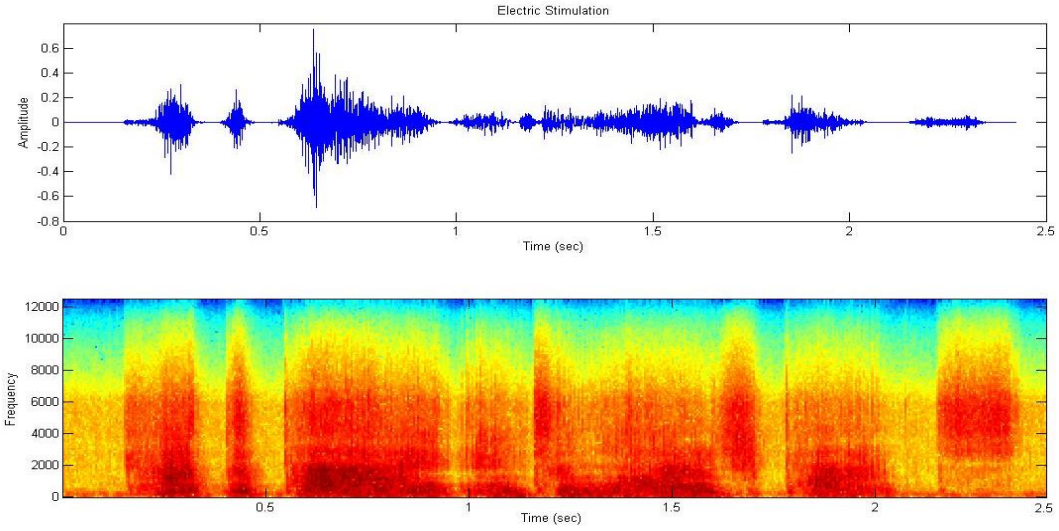


Figure 2.5. Effect of electric stimulation of the original speech spectrum of Figure 2.4.

2.8.3 Benefits of combined electric and acoustic stimulations (EAS)

Current cochlear implants provide very high level of speech understanding to hearing impaired listeners especially in quiet backgrounds. Despite great performance of these devices, not all

patients do well which is due to the limited ability of cochlear implants to provide adequate frequency resolution necessary for some complex listening situations. As depicted in the Figure 2.6, the combined electric and acoustic stimulations provide better representation of low-frequency and mid-to high frequency information for enhanced speech recognition.

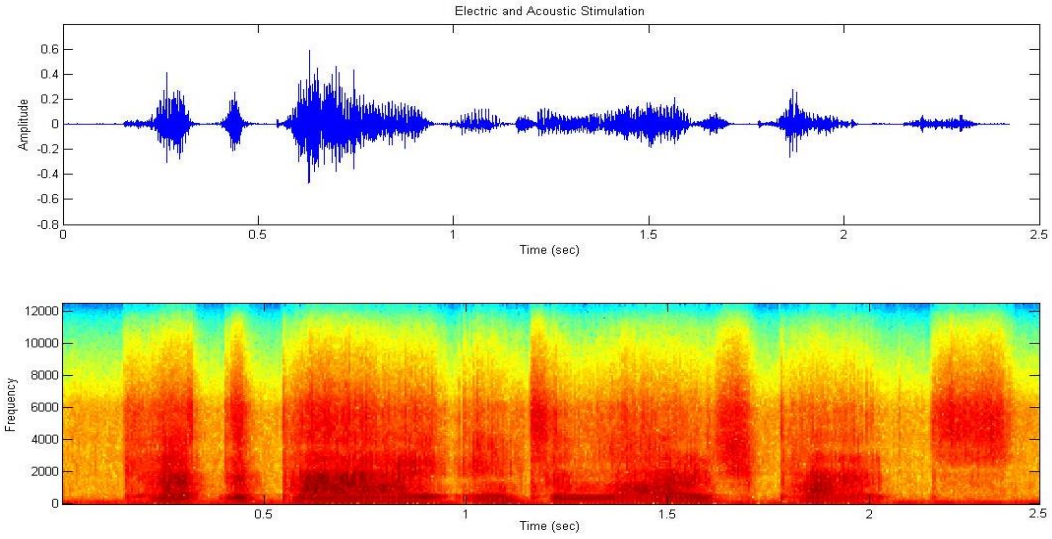


Figure 2.6. Effect of combined electric and acoustic stimulation of the original speech spectrum of Figure 2.4.

Von Ilberg *et al*. [107] reported the speech recognition in quiet results for one patient implanted with a long electrode inserted 20mm into the cochlea who had some preserved residual hearing 2 months following implantation, this patient's sentence recognition scores were 7% with the acoustic hearing alone, and increased to 56% with the addition of electric stimulation. Interestingly, sentence recognition with the CI alone was only 2%, suggesting a strong synergistic effect of the two modes of stimulation. Complex listening conditions with background noise require increased frequency resolution to extract the speech content from the mixed stimuli. In that case, increasing the number of channels could provide better frequency resolution. But due to the spatial interaction between the channels within the cochlea, even the best implant users could not

take advantage of more than 6–8 channels of electrical stimulation in noise [24]. Kiefer *et al.* [52] reported monosyllabic word understanding a group of 11 patients with preserved residual hearing. Mean acoustic-alone scores were 7%-correct, and the addition of electric stimulation increased this to over 60%, with several patients showing scores increasing to over 75%. Of these 11 patients, only 4 obtained A+E scores higher than the electric-alone score, suggesting that for many patients, the electric stimulation was providing the primary contribution to speech understanding.

To study the contributions of low-frequency acoustic information, simulated EAS experiments with normal-hearing subjects have been introduced to examine the mechanisms underlying the EAS benefits using controlled simulation conditions [9, 17, 68, 88, 98, 99]. As expected, all these studies reported synergistic EAS benefits for speech recognition in noise; more importantly, they also proposed several hypotheses about the beneficial low-frequency cues and the underlying mechanisms. Chang *et al.* [10] hypothesized that F0 present in low-frequency sound helped to segregate the target voice from competing voices and then group the speech information from temporal envelopes. EAS benefits were in part attributed to a better F0 representation [88] and glimpsing mechanism [62].

2.8.4 Other Benefits of EAS

Listening scores on tests for sentence understanding in noise were better in EAS conditions than in E conditions [17]. Many studies [17, 29, 43, 52, 59, 103, 107] have documented the improved speech intelligibility with combined electric and acoustic stimulations, especially in noise. A better representation could enable listeners to use voice pitch to separate a target voice from a background of other voices. This is an appealing hypothesis for the often reported improvement for speech understanding in noise in EAS conditions. The hypothesis is supported by several lines of evidence including those from simulations of EAS which show that low-passed speech

that contains only the fundamental frequency, and possibly a harmonic, of the voice aids speech understanding in noise [10].

In another study [59], it was found that low frequency acoustic hearing when combined with electric hearing significantly enhanced performance although the low-frequency acoustic hearing by itself produced no recognition for speech recognition in noise. However, melody recognition in the same group of subjects showed better performance with low-frequency acoustic hearing compared to electric hearing from cochlear implant. From these studies, it was concluded that combined acoustic and electric hearing provides better correlation between the salient pitch in low-frequency acoustic hearing and the weakly conveyed information in the envelope to enhance segregation between signal and noise for CI listeners.

CHAPTER 3

LITERATURE REVIEW

Cochlear implants allow hearing impaired listeners to achieve normal performance on speech recognition in quiet listening conditions. Despite the great efficacy of these devices, many cochlear implant listeners still struggle with perception of speech in adverse listening conditions. Several researchers have investigated the performance of cochlear implant devices under realistic conditions and proposed strategies to deal with the adverse impacts of the environment on the speech intelligibility by cochlear implant listeners.

3.1 Introduction

In this chapter, we provide a detailed review of research pertaining to enhancing the speech perception under three adverse environments identified namely telephone communication networks, reverberation and background noise.

In Section 3.2, the effects of reduced bandwidth of telephone speech on perception by cochlear implant listeners is discussed. Section 3.3 presents some of the bandwidth extension techniques proposed to enhance the perception of telephone processed speech. In Section 3.4, some of the impacts of reverberation on hearing ability of hearing impaired listeners are discussed in light of some of the prior studies. Some of the dereverberation algorithms are explored in Section 3.5. In Section 3.6, the effect of background noise on speech perception by cochlear implant listeners is discussed. Noise maskers such as speech-shaped noise and multi-talker babble are discussed in detail in this section. Some of the noise reduction methods developed to enhance the speech intelligibility will be presented in Section 3.7.

3.2 Effect of band-limiting on speech perception with cochlear implants

Rapidly changing technology requires fast information transmission. A telephone is a device that facilitate the transmission of information, whether official or personal. Individuals with severe to profound hearing loss find it difficult to use telephone for conversations. Telephones use the public switched telephone networks (PSTN). The bandwidth of the PSTN is determined by various speech components. The speech input for PSTN consists of sum of frequency components from 50 Hz to 7 kHz, but the speech output contains only frequency components between 300 Hz and 3400 Hz, thus compromising the quality to reduce the cost by transmitting limited frequency information. This bandwidth was selected as a result of speech intelligibility studies conducted at the Bell Labs by French and Steinberg [22]. Although this reduced bandwidth might not be an issue for normal hearing listeners, it has detrimental effect on listeners with hearing loss. To study the level of difficulty in understanding telephone speech, Kepler *et al.* [50] surveyed hearing impaired listeners with a questionnaire that investigates problems encountered by them while using the telephone. The results showed that 70% of the hearing impaired population reported problems using telephone and 75% suggested improvement in the telephone communications. The difficulty in understanding speech over the telephone was attributed to the limited frequency range, elimination of visual cues, and reduced audibility of telephone signals.

Frequency response of the communication channel has a severe impact on the performance of cochlear implants. Although cochlear implant listeners are capable of communicating over telephone, speech understanding is significantly worse than face-to-face conversations (wide-band speech). Perception of telephone speech was observed through speech tracking test and compared with the natural (non-telephone) speech [47]. Results indicated that the telephone speech communication ability of cochlear implant listeners was worse compared to natural speech.

Comparisons between normal hearing and cochlear implant listeners' performance for word discrimination of band-limited and wide-band speech were carried out by Milchard *et al.* [71]. Results showed that the normal hearing listeners did not experience any difficulty understanding speech. Word discrimination scores for telephone speech were 17.7% lower than scores for the wide-band telephone speech. Later, Fu and Galvin [26] measured vowel, consonant, and sentence recognition scores for simulated telephone speech and broadband speech and proved that there was no significant difference in vowel recognition scores between telephone and broadband speech. However, mean consonant and sentence recognition scores were significantly poorer with telephone speech. Cray *et al.* [13] investigated the use of telephone in cochlear implant listeners using the Clarion cochlear implant system manufactured by Advanced Bionics and found that 70% of the implanted population were telephone users out of which 30% used cellular phones for personal use. These findings attribute the advances in cochlear implant and telephone technology for the increased use of telephone by CI listeners.

3.3 Bandwidth extension techniques for telephone speech perception

In order to improve the ability of CI listeners to perceive the telephone speech, several researchers had suggested a change in the public switched telephone network (PSTN) to transmit the wide-band speech. Several wideband speech coding schemes have been developed for the increased acoustic bandwidth (50 Hz – 7000 Hz). Some of the more economical approaches to bandwidth extension are discussed below.

3.3.1 Frequency selective amplification and compression method

Terry *et al.* [101] developed two signal processing strategies to compensate for the high frequency hearing loss in hearing impaired listeners. The two strategies were: Frequency domain and Time-domain processing. Fast Fourier Transform was used to extract the short-time spectrum which was

modified. Time domain processing involved passing the signal through a bank of finite impulse response filters.

Speech signals were transformed in frequency domain, by evaluating the spectrum of the signal. In this method, the transformation was obtained by multiplying the speech spectrum by a frequency function. It was necessary to convert the consecutive samples of the signals to and from the frequency domain. Since too short of the window length will result in poor spectra, and too long of the window length will result in smearing of the temporal modulations, consideration was given to the number of samples or window length that was used to estimate the short-term spectra of the speech signals. In the second method, speech signals were transformed into time domain directly by convolving the signal against the other time functions. Signals were digitally filtered using single or multiple finite impulse response (FIR) functions. Both frequency and time domain methods were examined to achieve frequency selective amplification and compression.

Frequency domain method (Short-term spectrum modification): In this method the signal was multiplied by a hamming window of size equal to the size of the FFT, which was chosen to be 256 samples. Each sample was Fourier transformed and multiplied with a frequency function. The modified signals was then inverse transformed to reconstruct the time signal. Overlap and add method was used to reconstruct the time signal. The time signal was sampled using 16 A/D converter and the spectrum was divided into six bands. Two of these bands were outside the telephone speech bandwidth. Four channels were selected within the telephone bandwidth (321.5 to 500Hz, 562.5 to 1000 Hz, 1062.5 to 2000Hz and 2062.5 to 3000Hz) for which independent gains and compressions were applied. The gain factor was related to the hearing loss in that channel. If amplitude compression was used then the gain factor was modified by the compression ratio of the channel.

Time domain Method (FIR Digital Filter Bank): In this method, hamming window was used to construct a window function technique for FIR filter and the input signals are processed using a bank of variable FIR filters. The center frequencies and bandwidths were chosen similar to the frequency domain method, and then all four filter outputs were mixed. If the gain of the filter was set to one then the frequency response across the telephone bandwidth was essentially flat. The gain factors are determined by the audiometric thresholds to compensate for the individual hearing loss.

There were three conditions tested for both types of signal processing: Frequency shaping (FS), Frequency shaping plus compression (FSC), and No shaping or compression (NSC). No shaping or compression was a control condition where frequency shaping was flat and no compression was applied. In frequency shaping (FS) condition, mild hearing losses were mirrored by the frequency-selective amplification. For a loss greater than 40dB, the gain of at the frequency was limited to 40dB. In frequency shaping plus compression, the compression ratio was determined by the difference between the hearing threshold and the uncomfortable loudness level measured at an audiometric frequency within the band.

They tested 16 subjects with sensorineural hearing loss and each condition was presented at 3 levels 10dB, 20dB and 30dB above the average hearing thresholds at 1000Hz and 2000Hz. Results indicated no significant difference between time domain and frequency domain processing. For both methods, frequency shaping condition showed significant effect. Frequency shaping improved the intelligibility in FS and FSC conditions. In another experiment, subjects were asked to listen to a female speaker and then have a conversation with the female and male talker over the telephone. 13 out of 16 subjects preferred FS and FSC signal over unshaped telephone speech. In conclusion to their work, the authors believe that frequency shaping would

increase the intelligibility at least in mild hearing loss. The concerning factors of these adaptive processing are the processing delays and the dependence on the audiometric frequency information.

3.3.2 Telephone adapter approach

Ito, Nakatake and Fujita [47] investigated the hearing ability of cochlear implant listeners using telephone adapters. These devices were initially used to reduce the noise levels in the telephone speech and to record the voice into a tape recorder. The telephone adapter used in their study consisted of a parallel resonant circuit tuned as a band-pass filter to work as a matched filter for telephone speech. The characteristics of the adapter are to cover the telephone band with a frequency component between 300 Hz and 3.5 kHz and to attenuate as much as 20 dB at the both edges of the pass band, eliminating half of the noise power while maintaining the essential speech components. Due to the advantages of spectral peak (SPEAK) stimulation mode over the multipeak (MPEAK) mode, the authors were interested in testing the usefulness of telephone adapter with this mode.

In their study, 10 postlingually deafened adult cochlear implant listeners using SPEAK coding strategy were tested for vowel-confusion, consonant-confusion and speech-tracking. Results of the vowel-confusion test; the average percent correct scores for natural voice and for the telephone adapter voice were 96.8% and 94.8%, respectively. For consonant-confusion test; the average percent correct scores for natural voice and for the telephone adapter voice were 56.4% and 48.6%, respectively. And the results of the speech-tracking test; for natural voice, for the telephone voice, and for the telephone adapter voice were 111.5, 62.4 and 109.3 phrases per 5 minutes, respectively.

3.3.3 Bandwidth extension technique based on Hidden Markov model (HMM)

Peter Jax and Peter Vary developed a bandwidth extension algorithm [81].The first step in their bandwidth extension algorithm was the estimation of the spectral envelope of the original wideband speech signal. The idea was to estimate the spectral envelopes of the extension band represented by the cepstral vector $\tilde{y}_{eb}$. The estimation is based on the observation of a feature vector x that is extracted from the narrowband speech signal $s_{nb}(k)$. The true spectral envelope in the extension band is obtained as follows: For the training process and later on for the evaluation of the overall system, the true wideband signal was split into two sub-band signals which contain the narrowband components $s_{nb}(k)$ and the extended frequency components $s_{eb}(k)$ respectively. An auto regressive model is fitted to each frame (20ms) of the extension band signal $s_{eb}(k)$ which was accomplished by a conventional LPC analysis. The spectral envelope of the extension band is described by the coefficient set a_{eb}= [a_{eb}(1)… a_{eb}(Na)] of the resulting all-pole filter and the corresponding gain factor σ_{eb}

$$E_{eb}(e^{-j\Omega n}) = \frac{\sigma_{eb}^2}{\left|A_{eb}\,(e^{j\Omega})\right|^2} \tag{3.1}$$

Using the definition of the real cepstrum:

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} ln \left|\frac{\sigma_{rel}^2}{\left|A_{eb}\,(e^{j\Omega})\right|^2}\right| e^{-j\Omega n}\, d\Omega \tag{3.2}$$

Using Fourier representation, the envelope of the extension band can be approximated as,

$$\sum_{n=1-d}^{d-1} c(n)\, e^{-j\Omega n} \approx ln\frac{E_{eb}}{\sigma_{nb}^2} = \ln\frac{\sigma_{rel}^2}{\left|A_{eb}\,(e^{j\Omega})\right|^2} \tag{3.3}$$

with $\sigma_{rel} = \frac{\sigma_{eb}}{\sigma_{nb}}$. The cepstral coefficient $c(n)$ is determined from the AR coefficient a_{eb}. The representation of the extension band spectral envelope is finally determined by the weighted cepstral coefficient yielding d-dimensional vectors $y_{eb} = [y_{eb}(0), y_{eb}(1), \ldots \ldots y_{eb}(d-1)\ ]$

$$y_{eb}(n) = \begin{cases} \frac{1}{\sqrt{2}}\, c(n) & ; \quad if\ n = 0 \\ c(n) & ; \quad if\ 1 \leq n \leq d \end{cases} \tag{3.4}$$

The estimate $\tilde{y}_{eb}$ of the envelope of the extension frequency band was determined in the application phase of the algorithm. The vector estimation $\tilde{y} = \tilde{y}_{eb}$ representing the extension band spectral envelope was based on an HMM of the speech generation process. Each state S_i of the HMM (i =1, 2…., NS) was assigned to a typical speech sound (frame of 20ms) which were associated with a representative envelope $\hat{y}_i$ in the extension band. The states of the HMM are defined by a vector quantization of the spectral envelope representation y. Each state S_i of the HMM corresponds to one entry $\hat{y}_i$of the vector quantizer codebook C = { $\hat{y}$1, $\hat{y}$2…. , $\hat{y}$NS} such that the number of states in the HMM is the same as the number of entries in the codebook. However, wideband speech S_{wb} was available only in the training phase, whereas in the application phase of the bandwidth extension algorithm the states S_i had to be identified from the frames of the narrowband speech signal S_{nb} by classification. Therefore, an estimator had to be developed which extracts from the narrowband speech relevant information about the spectral envelope in the extension band. The estimate $\tilde{y}_{eb}$ was combined with the narrowband signal frame in a short-term power spectrum domain. The auto-correlation function uses standard linear prediction analysis to determine the AR coefficient $\tilde{a}_{wb}$. An estimate of the narrowband excitation signal $\tilde{u}_{nb}(k)$ was derived by applying the FIR analysis filter $\tilde{A}(z)$ to the narrowband input signal $s_{nb}(k)$.The extension of the excitation signal converts the narrowband excitation signal $\tilde{u}_{nb}(k)$

into an extended version $\tilde{u}_{wb}(k)$ by exploiting the spectral flatness. The estimated wideband excitation signal $\tilde{u}_{wb}(k)$ was then fed into the wideband all-pole synthesis filter 1/A(z) to synthesize the enhanced output speech $\tilde{s}_{nb}(k)$.

The principal part of the algorithm is the human vocal tract, which analyzes and extends the envelope of the frequency spectrum of the speech signal. To minimize the minimum mean square error, power spectrum estimation was done using weighted cepstral domain. The minimization of the MSE corresponds to the explicit optimization of the perceptually relevant mean LSD measure. In listening tests, the additional utilization of the HMM yields a significant reduction of unnatural artifacts in the enhanced speech. In their work, the actual estimation procedure is based on an HMM of the signal source. By taking into account the statistics of the HMM state sequence, their MMSE estimation rule additionally considers the observations from previous signal frames.

3.3.4 Bandwidth extension technique based on Gaussian Mixture Model (GMM)

Recently Liu, Fu, and Narayanan [63] developed a bandwidth extension techniques based on Gaussian mixture models that partly restore higher frequency information. Gaussian mixture model (GMM) was used to model the spectrum distribution of narrow-band speech. The relationship between wide-band and narrow-band speech was used to recover the missing information based on the available telephone band speech which was learned a priori in a data driven fashion. Their technique was divided into three parts: GMM based spectral envelope extension, excitation spectrum extension and speech analysis and synthesis.

GMM based spectral envelope extension: A GMM represents the distribution of the observed parameters by *m* mixture Gaussian components in the form of

$$p(x) = \sum_{i=1}^{m} \alpha_i \, N(\mathrm{x}, \mu_i, \Sigma_\mathrm{i}) \tag{3.5}$$

where α_i denotes the prior probability of component *i* ($\sum_{i=1}^{m} \alpha_i = 1 \; and \; \alpha_i \geq 0$) and $N(\mathrm{x}, \mu_i, \Sigma_\mathrm{i})$ denotes the normal distribution of the *i*th component with mean vector $\mu_i$and covariance matrix Σ_i in the form of:

$$N(\mathrm{x}, \mu_i, \Sigma_\mathrm{i}) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp[-\frac{1}{2} (\mathrm{x} - \mu_i)^T \sum_i^{-1} (\mathrm{x} - \mu_i) \tag{3.6}$$

where *p* is the vector dimension. The parameters of the model (α, μ, Σ) were estimated using the well-known expectation maximization algorithm. Consider **x** = [**x**1 **x**2…. **x**n] to be the sequence of *n* spectral vectors produced by the narrow-band telephone speech, and **y** = [**y**1 **y**2….**y**n] to be the time-aligned spectral vectors produced by the wide-band speech. The objective of the bandwidth-extension method was to define a conversion function $F(\mathrm{x}_t)$ such that the total conversion error of spectral vectors:

$$\varepsilon = \sum_{t=1}^{n} (y_t - F(\mathrm{x}_t))^2 \tag{3.7}$$

was minimized over the entire training spectral feature set, using the trained GMM that represents the feature distribution of the telephone speech. A minimum mean square error method was used to estimate the conversion function. The conversion function was given by:

$$F(\mathrm{x}_t) = \sum_{i=1}^{m} P(\mathrm{C}_i|\mathrm{x}_t) [\mathrm{v}_i + \mathrm{T}_i \sum_i^{-1} (\mathrm{x}_t - \mu_i)] \tag{3.8}$$

where $P(\mathrm{C}_i|\mathrm{x}_t)$ is the posterior probability that the *i*th Gaussian component generates x_t ; v_i and T*i* are the mean wide-band spectral vector and the cross-covariance matrix of the wide-band and narrow-band spectral vectors, respectively. When a diagonal conversion is used (i.e., T*i* and Σ_i are diagonal), the above optimization problem simplifies into a scalar optimization problem, and the computation cost is greatly decreased.

Excitation spectrum extension: Two methods considered for excitation spectrum extension in their study. These were spectral folding and spectral translation. Spectral folding simply generates a mirror image of the narrow-band spectrum for high-band spectrum. In spectral translation, the excitation spectrum of the narrowband speech, obtained from Fourier transformation of the time domain signal, is translated to the high-frequency part and padded to fill the desired whole band. A low pass filter is applied to do spectral whitening, such that the discontinuities between the translations are smoothed.

Speech analysis and synthesis: They extracted the Mel-scaled line spectral frequency (LSF) features (18th order) and energy to model the spectral characteristics of speech in a 19 dimensional space. The spectral features between narrow-band and wide-band speech were aligned with dynamic time warping computation. The spectral mapping function between narrow-band and wide-band speech was trained with 200 randomly selected sentences from the IEEE database (100 sentences from a female talker and the other 100 sentences from a male talker). The excitation component between 1 and 3 kHz was used to construct the high-band excitation component because the spectrum in this range was relatively white. A low pass Butterworth filter (first order with cutoff frequency 3000 Hz) was used to do spectral whitening. The synthesized high-band speech (i.e., frequency information above 3400 Hz) was obtained from high pass filtering the convolution result of the extended excitation and extended spectrum. It was then appended to the original telephone speech to render the reconstructed wide-band speech that covered the frequency band from 300 to 8000 Hz.

Seven CI subjects (two women and five men) participated in their study. Some subjects used ACE and others SPEAK strategy with 8 electrode respectively. The recognition score with

the bandwidth-extension method was about 3.5% higher than without the bandwidth-extension method. The improvement was small but significant (paired t-test, $p = 0.050$).

3.3.5 Conclusion

In the above section, we discussed bandwidth extension techniques developed to improve the recognition of telephone-processed speech by hearing impaired listeners. To summarize the literature, frequency selective amplification and compression method was used to compensate for the high frequency information and significant improvement in telephone speech intelligibility for the hearing impaired. This study required audiometric frequency data from users for best performance and suffers processing delays. Another approach with telephone adapters was used to improve telephone speech intelligibility by CI users. However, these instruments are not easily available and therefore has limitations on implementation. Recently, a technique based on GMM model was utilized to partly restore high-frequency information for telephone speech, although the speech synthesis could not completely eliminate the perceptual distortions. These studies have not been able to distinguish between the effects of low frequency and high frequency information and does not study the combined electric and acoustic stimulations in cochlear implant listeners.

3.4 Effects of reverberation on speech perception

Reverberation is a phenomenon of acoustical enclosures where impulsive sound signals from the source get reflected by surroundings and arrive at the receiver. Reverberation can significantly impact the speech perception and intelligibility. Many researchers who studied the effects of reverberation have noticed flattening of formant transitions in vowels, energy smearing within the phoneme, reduced temporal amplitude modulations, and increased low-frequency energy which masks high-frequency components [3, 6, 7, 57, 74, 75, 76, 77].

3.4.1 Self-masking and Overlap-masking Effects

In acoustic reverberant environments, the received signal generally consists of a) Direct sound that travels from the sound source directly to a listener; b) Early reflections, which arrives at the listener within approximately the first 50ms after the direct sound; and c) Late reflections, which arrive at the listener with a much longer delay after the arrival of the direct sound. The first sound that is received through free-field, i.e., without reflection, is the direct sound. Sometime later the sounds which were reflected off one or more surfaces (walls, floor, furniture, etc.) will be received. These reflected sounds are separated in both time and direction from the direct sound. Early reflections will vary as the source or the microphone moves within the space, and gives us information about the size of the space and the position of the source in the space. Early reflections are not perceived as a separate sound to the direct sound so long as the delay of the reflections does not exceed a limit of approximately 50ms with respect to the arrival time of the direct sound. The direct sound travels from the sound source directly to a listener and is, generally, not influenced by the properties of the enclosure. Therefore, these are considered beneficial to the overall speech intelligibility, as they can be perceptually integrated with the direct sound [8, 45, 60]. Late reflections, on the other hand, consist of a dense succession of echoes with diminishing intensity and typically arrive at the listener with a much longer delay after the arrival of the direct sound component. Late reflections tend to fill the gaps in the temporal envelope of speech and reduce the low-frequency envelope modulations important for speech intelligibility. These late reflections have been shown to produce mainly overlap-masking effects due to the temporal overlap of reverberant energy from a preceding phoneme onto the following phoneme [15, 31, 45, 55]. In contrast, early reflections are often associated with moderate self-masking (coloration)

effects due to the internal smearing of energy within each phoneme and thus early reflected sound not only improve the signal-to-noise ratio but also modify the sound quality in a pleasant way [55].

Bolt and Mac Donald [7] initially studied the theory of speech masking in reverberated signals and computed the articulation index. They hypothesized that most of the information is carried in the leading edge of the pulse and self-masking was caused by energy components of leading portion on the remaining portion. Later it was found that this speech degradation is due to two different but interdependent effects: (1) overlap-masking and (2) self-masking effects. Overlap-masking is due to the overlap of reverberant energy of a preceding phoneme on the following phoneme. The additive reverberant energy fills in the gaps and silent intervals (e.g., stop closures) associated with vocal tract closures. This effect is more evident for low-energy consonants preceded by high-energy voiced segments. Self-masking is caused by the internal smearing of energy within each phoneme. It is substantially smaller when compared to the overlap-masking of consonants in the final position [15, 75, 76]. This effect is particularly evident in reverberant sonorant sounds (e.g., vowels), where the formant transitions become flattened [55].

3.4.2 Reverberation Time (RT_{60})

Reverberation time is another term used to define time for the sound to die away after the sound source ceases, but that of course depends upon the intensity of the sound. When the sound source stops, the reverberant sound level begins to fall, but it takes some time for it to become inaudible. The time taken for the level to fall by 60 dB is known as the reverberation time (RT_{60}). The reverberation time (RT_{60}) is estimated as the interval in which the reverberating sound intensity, due to decaying reflections, reaches one millionth of its initial value. In other words, it is the time it takes for the reverberation level to drop by 60 dB below the original sound energy present in the room at a given instant. Therefore, designing acoustics of a room becomes increasingly important

for creating a universally accessible environment for speech communication especially for hearing impaired listeners. Many investigators have reported that for normal-hearing (NH) listeners, individuals suffering from sensorineural hearing impairment, and cochlear implant (CI) listeners, the detrimental effects of reverberation on speech perception exhibit a common pattern. In all of these populations, the intelligibility of reverberated speech progressively declines with a linear increase in reverberation time, RT_{60} [15, 33, 45, 54, 55, 70, 74, 75, 76, 77, 78].

Although overall speech recognition by normal-hearing listeners may not be degraded until the reverberation time (RT_{60}) exceeds approximately 1.0 s [54, 76], speech intelligibility measured in listeners with sensorineural hearing loss has been shown to deteriorate considerably even in situations where the reverberation time exceeds 0.5 s. A study on vowel recognition that assessed the performance of ten elderly adults with binaural sensorineural hearing loss showed that the mean vowel recognition score in reverberation time of 1.2 s was approximately 12 percentage points lower than the mean score obtained in a non-reverberant (anechoic) listening condition [75].

3.4.3 Effect of reverberation on speech perception with cochlear implants

Compared to the literature on the effects of reverberation on speech perception by normal hearing (NH) listeners and listeners with sensorineural hearing loss, far fewer studies have examined the intelligibility of reverberated speech perceived by CI listeners. Reverberation greatly reduces the modulations present in the envelopes making it extremely challenging for CI users to extract useful information about speech (e.g. F0 modulations, location of syllable/word boundaries). Poissant *et al.* [85] and Whitmal and Poissant [108] were among the first to investigate speech perception in CI listeners in reverberant environment. They used reverberant stimuli processed into 6-24 channels using tone-excited vocoders and tested the normal hearing listeners. It was shown that the percent correct recognition scores will deteriorate significantly when a small number of

spectral bands are used to vocode speech signals inside a simulated reverberant field. Their results indicated low performance with 6 channels and mild effect of reverberation time when the reverberant stimuli were vocoded into 12 or more channels.

A channel specific model to detect reverberation in CI was developed by Desmond *et al.* [14]. This study models speech stimuli in a cochlear implant on a per-channel basis both in quiet and in reverberation, and assesses the efficacy of these models for detecting the presence of reverberation. This study was able to successfully detect reverberation in cochlear implant pulse trains, and the results appear to be robust to varying room conditions and cochlear implant stimulation parameters. Reverberant signals were detected 100% of the time for a long reverberation time of 1.2 s, and 86% of the time for a shorter reverberation time of 0.5 s.

From the literature summarized above, it quickly becomes evident that the effects of additive reverberant energy on speech identification by both normal-hearing and hearing-impaired listeners can be quiet detrimental. However, to-date, very little is known about the extent to which reverberation can be suppressed to improve the listening abilities of CI recipients.

3.5 Reverberation suppression strategies

Addressing the impacts of reverberation has recently become an area of intense research activity and has led to several dereverberation algorithms for speech enhancement [34, 55, 73]. Reducing distortion due to additive reverberation through inverse filtering is one of the first and remains one of the most commonly used methods today [73]. In most methods, the main idea of reverberation cancellation is to pass the reverberant (corrupted) signal through a finite impulse response (FIR) filter that inverts the reverberation process, thus recovering the original signal. There are channel-selection criterion based methods [34, 55] that were recently developed to tackle the reverberation.

In the next section we discuss some of the dereverberation algorithms developed to reduce the detrimental impact of reverberation on CI listeners.

3.5.1 Inverse Filtering Technique

The main idea of reverberation cancellation or speech de-reverberation in this approach was to apply multiple-input/output inverse theorem (MINT) [73]. An acoustic system consisting of the speaker S1 and the microphone (M) has a transfer function denoted by G (z^{-1}) which represents the direct sounds and reflective sounds between the source and the receiver. Here, an acoustic system was considered to be a multiple input (or multiple output) linear impulse response (FIR) system and hence the approach is referred to as MINT. To explain the method in detail, they used single input and single output linear FIR system. When the filter is the inverse of the system, it satisfies:

$$d\,(k) = \; g(k) \circledast h(k) \tag{3.9}$$

Where

$$d(k) = \begin{cases} 1, & k = 0; \\ 0, & k = 1,2,3, \dots \dots \dots \end{cases} \tag{3.10}$$

As many systems in room acoustics can be modified to multiple input linear systems, the exact inverse can be constructed. With two signal transmission channels, the FIR filters, $H_1(z^{-1})$and $H_2(z^{-1})$ satisfies the following:

$$D(z^{-1}) = \; G_1(z^{-1})\, H_1(z^{-1}) + \; G_2(z^{-1})\, H_2(z^{-1}) \tag{3.11}$$

There exists a pair of FIR filters, $H_1(z^{-1})$and $H_2(z^{-1})$, that can realize exact inverse filtering of a two input single-output linear FIR system by connecting these to the inputs of S_1 and S_2, respectively. Therefore, it becomes possible to reproduce the desired acoustic signals at the microphone (M) without any distortion caused by wall reflections using the proposed principle.

The system's two signal-transmission channels are denoted as $U_1(z^{-1})$ and $U_2(z^{-1})$. Two FIR filters $V_1(z^{-1})$and $V_2(z^{-1})$ are assumed to be connected to the outputs of $U_1(z^{-1})$ and $U_2(z^{-1})$, respectively. To reconstruct the input signal of the system, $V_1(z^{-1})$and $V_2(z^{-1})$ must satisfy the expression:

$$1 = \ U_1(z^{-1})\, V_1(z^{-1}) + \ U_2(z^{-1})\, V_2(z^{-1}) \tag{3.12}$$

Here, the above-mentioned principle is extended for inverting a multiple-input multiple-output linear FIR system. Multiple finite-impulse response (FIR) filters (transversal filters) are used to construct the inverse by adding some extra acoustic signal-transmission channels produced by multiple loudspeakers or microphones. The coefficients of these FIR filters can be computed by the well-known rules of matrix algebra.

To verify the applicability of the method to a sound field, an inverse-filtering experiment was conducted in the frequency band 315-3150 Hz. It was shown that the proposed method is greatly superior to previous methods that use only one acoustic signal-transmission channel. The results prove the possibility of sound reproduction and sound reception without any distortion caused by reflected signals. However, the main drawback of inverse filtering approaches is that the acoustic impulse response must be known in advance or alternatively needs to be "blindly" estimated for successful de-reverberation. It was known to be a fairly difficult and computationally expensive task. The error in the proposed method was thought to be due to the accuracy limits of the digital computer used in the experiment.

3.5.2 Ideal reverberant mask (IRM) approach

An Ideal reverberant mask (IRM) is a binary mask for reverberation suppression which was computed using signal-to-reverberant ratio (SRR). This channel selection criterion produced substantial intelligibility gains for CI users even in highly reverberant environments [55].

Maximum selection criterion adopted in ACE coding strategy erroneously picks amplitudes during the gaps present in most unvoiced segments of the utterance. As a result, the vowel and consonant boundaries are smeared, making it difficult for the listeners to use effectively lexical segmentation cues needed for word retrieval. The ideal condition is to select the amplitudes corresponding only to the direct sound and early reflections, and at the same time discard the amplitudes corresponding to the late reflections. Such a criterion, however, would require access to the acoustic impulse responses, which in practical scenarios may not be available. So this strategy presents a selection criterion based on the SRR, which for each channel is computed as follows:

$$SRR(t,k) = 10log_{10}\ \frac{|X(t,k)|^2}{|Y(t,k)|^2} \tag{3.13}$$

where *X(t, k)* and *Y(t, k)* denote the clean and reverberant signals, respectively, *t* corresponds to the time-frame index, and *k* defines the frequency or channel index. If SRR value is large, it would mean that the energy from the direct signal (and early reflections) dominates, as is often the case during the voiced segments (e.g., vowels). In contrast, if SRR value is small, it would mean that the reverberant energy, composed of the sum of the energies from the early and late reflections, dominates. This is caused primarily by overlap-masking occurring primarily during the gaps. By comparing the individual channel-specific SRR values against an empirically determined threshold value, T. They were able to remove the reverberant energies in the gaps, which enabled them to minimize the overlap-masking effects.

The SRR selection criterion was implemented by multiplying the reverberant signal by a binary time–frequency (T-F) mask or equivalently a binary gain function. This mask could take the value of 1 when SRR > T and zero otherwise. The dereverberated signal at T-F (t, k) is obtained as follows:

$$\hat{X}_{DE}(t,k) = Y(t,k).IRM\ (t,k) \tag{3.14}$$

where $\hat{X}_{DE}(t,k)$ denotes the estimated dereverberated signal and *Y(t, k)* is the reverberant signal. The ideal reverberant mask, *IRM (t, k)* is given by

$$IRM\ (t,k) = \begin{cases} 1, & \text{SRR}(t,k) > T' \\ 0, & \text{Otherwise} \end{cases} \tag{3.15}$$

where T' represents the threshold value, expressed in dB. We refer to *IRM* as the ideal reverberant mask because its construction requires prior knowledge of the original (uncorrupted) acoustic information. It is worth mentioning that different forms of binary T-F masks have been previously used in other applications to suppress noise. These T-F masks were based on the local SNR criterion rather than the SRR criterion.

Speech was first synthesized using the proposed channel-selection criterion and then fed as input to the SPEAR3 speech processor. They used 128-channel (N = 128) fourth-order gammatone filter-bank, with center frequencies equally spaced on the equivalent rectangular bandwidth (ERB) scale covering a frequency range between 50 and 8 kHz to derive the T-F representation of the clean speech and the reverberant inputs. The filtered waveforms were then divided into 20ms frames with 50% overlap between successive frames, and the short-time energies of the filtered waveforms are computed. They compared the energy of the clean and reverberant signals by calculating the SRR independently in each individual T-F channel. The resulting SRR for each T-F unit was then compared against a preset threshold value T to determine whether to retain a specific T-F unit or to discard it. Out of the 128 initial filtered waveforms, only

the T-F units where the energy of the clean signal exceeds that of the reverberant signal by the specified threshold value, such that SRR(t, k) > T, are retained.

Head related transfer functions from Rychtarikova *et al.* [89] were used to simulate reverberant conditions. To generate the stimuli used in our study, the HRTFs obtained for each reverberation condition were convolved with the speech files from the IEEE test materials using standardized linear convolution algorithms in MATLAB. The sound pressure level (SPL) measured at the center of the artificial head was fixed at 70 dB SPL. The overall reverberant characteristics of the experimental room were altered by adding floor carpeting and absorptive panels on the walls and the ceiling, as described in [89]. The results indicated that in a highly reverberant scenario, the proposed strategy led to substantial gains (over 60 percentage points) in speech intelligibility over the subjects' daily strategy. Further analysis indicated that the proposed channel-selection criterion reduces the temporal envelope smearing effects introduced by reverberation and also diminishes the self-masking effects responsible for flattened formants. The problem with this strategy was that the construction of the SRR criterion assumed a priori knowledge of the clean target envelopes.

3.5.3 Blind reverberant mask (BRM) approach

Hazrati *et al*. [34] were motivated by the results obtained with IRM strategy and proposed to extend it to a blind channel-selection criterion for dereverberation. Their new channel-selection strategy neither required information of clean (anechoic) signal nor the prior knowledge of the room impulse response. They used linear prediction analysis to obtain a residual signal, which helps in determining the residual-to-reverberant ratio (RRR) of individual frequency channels. This ratio was used as a channel-selection criterion, where channels with RRR less than an adaptive threshold were retained while the rest were zeroed out for each frame.

This algorithm was divided into four stages. The first stage computes the time-frequency representation of the speech signal by passing it through a set of bandpass filters and blocking the bandpass filtered outputs to short time overlapping frames, where each short time frame at each frequency bin corresponds to a T-F unit. In order to make decisions for classifying the T-F units as speech or (late) reverberation dominant, features are extracted in the second stage for all T-F units. The features are then passed into the next stage where the threshold value for each unit is computed, the T-F units are classified as speech or reverberation dominant, and their corresponding mask value is set to 1 or 0 accordingly. This binary mask provides an estimate of the IRM (ideal case) and is applied to the T-F representation of the reverberant signal in the last stage. The binary-masked bandpass filtered reverberant speech signals are summed across different frequency bins to re-synthesize the dereverberated speech. The input reverberant speech, r(n), is passed through J-channel (here, J is set to 64) Gammatone filterbank, with quasi-logarithmically spaced center frequencies and short-time frame blocking to decompose into T-F units. The T-F decomposed signal is denoted by r(t, j) where t and j represent time frame and frequency band indices, respectively. A discriminative feature is computed to identify the peaks and valleys in each band using a ratio of the variance of the signal raised to a power and the variance of the absolute value of the signal defined as follows:

$$f_M(t,k) = 10log_{10}\ \frac{\sigma^2_{r'}(t,j)}{\sigma^2_{|r|}(t,j)} \tag{3.16}$$

Where $r'(t,j) = |r(t,j)|^{\alpha}$, and $|r(t,j)|$ is the absolute value of the *L (*frame size) dimensional reverberant vector in frame *t*, and frequency band j. The parameter α is set to 2.1 experimentally. In order to make decision on the features extracted using (3.16), as to whether they are reverberation-dominant or speech-dominant, they used binary mask threshold estimation

technique. The input to this histogram-based threshold estimation technique at time frame *t* and frequency band *j* is the following feature vector containing features of '*Lp*' previous and '*Lf*' future frames:

$$f_{hist}(t,k) = \{f_M(t-Lp,k), \ldots\ldots f_M(t+Lf,k)\} \tag{3.17}$$

They defined between-class variance with distinct intensity level index, *tr*, as:

$$\sigma_B^2(tr) = \frac{(m_G\,(P_S\,(tr) - m(tr))^2}{P_S\,(tr)\,(\,1 - \,P_S\,(tr))} \tag{3.18}$$

Where m_G is the global intensity mean, $m(tr)$ is the cumulative mean, and $P_S(tr)$ is the cumulative sum defined as follows:

$$m_G(Tr) = \sum_{i=1}^{Tr} i\,.\,p_i\,; \quad m(tr) = \sum_{i=1}^{tr} i\,.\,p_i\,; \quad P_S(tr) = \sum_{i=1}^{tr} p_i \tag{3.19}$$

where p_i denotes the normalized histogram of the feature vector in (3.17).The equation (3.18) is used to find the optimum threshold level *tr** in the following manner:

$$\sigma_B^2(tr*) = \max_{tr=1,\ldots\ldots Tr} \sigma_B^2(tr) \tag{3.20}$$

where *Tr* is the total number of distinct levels of the histogram of the input feature vector (*fhist*). This algorithm will compute inaccurate threshold levels resulting in incorrect decisions, if the long-term windowed feature vector contains only silence. Therefore, a minimum threshold level (*tr0*) is set to discriminate silence from speech. Use of the long-term windowed feature vectors along with *tr0* results in a robust and effective adaptive threshold level estimation. If the feature value for a T-F unit is greater than the adaptive threshold of that specific T-F unit, the frame is classified as reverberation-free, otherwise it is considered as reverberation-dominant. Frames

classified as reverberation free are retained, while reverberation-dominant frames are zeroed out. This forms a binary blind reverberant mask, *BRM (t, j)* which is defined as:

$$BRM\ (t,j)\ = \begin{cases} 1, & f_m(t,k) > \max(\ tr^*\ (t,j), tr0) \\ 0, & \text{Otherwise} \end{cases} \tag{3.21}$$

where $f_m(t,k)$ is the feature extracted in (3.16). The enhanced signal is obtained after applying the binary mask, estimated based on comparing features and threshold levels to the reverberant signal. This technique removes the reverberation-dominant T-F units resulting in restoration of the word/syllable boundaries.

They compared the BRM results with the IRM as it provides the upper bound in performance. Subjective listening tests were conducted with six CI listeners using Nucleus processors in simulated rooms for three reverberant conditions of RT_{60} = 0.3s, 0.6s and 0.8s. Performance was measured in terms of speech intelligibility and for *BRM* criterion, the subjective intelligibility scores improved by 2.9%, 23.7%, and 27.0% absolute percentage points for RT_{60} = 0.3, 0.6, and 0.8 s conditions, respectively. These improvements are found to be statistically significant ($p < 0.05$) at RT_{60} = 0.6 and 0.8 s. Like IRM method, BRM also removes the overlap masking effect of reverberation; however, it is clear from the comparison of this method to IRM, that the BRM makes mistakes in low frequency regions which is one of the main reasons for the intelligibility gap between IRM-processed and BRM-processed signals.

3.5.4 Conclusion

In the above section, we discussed dereverberation algorithms that has been developed to suppress the reverberation and improve the intelligibility of reverberated speech in cochlear implant listeners. To summarize the literature, inverse filtering was among the initial methods developed

to tackle reverberation. For application to cochlear implants, this method used multiple finite-impulse response (FIR) filters to construct the inverse by adding some extra acoustic signal-transmission channels produced by multiple loudspeakers or microphones. Although this method was superior to previous methods in inverting the process of reverberation and produced improved speech quality without any distortion, the acoustic impulse response must be known in advance or estimated "blindly" for successful de-reverberation, which was the main drawback of this method.

A channel selection strategy was developed which gained popularity for dereverberation, ideal binary masking. A selection criterion was designed based on signal-to-reverberant ratio (SRR), which was computed by multiplying the reverberated signal with the binary mask. By comparing the individual channel-specific ratios against an empirically determined threshold value, T, they were able to remove the reverberant energies. The only problem in this approach was the prior knowledge of clean signal to compute the signal-to-reverberant ratio. Soon after, a blind channel selection approach was introduced, which was independent of the information from clean signal. A residual-to-reverberant ratio (RRR) was computed using a linear prediction analysis of the reverberant signal and was used as a channel-selection criterion. The channels with RRR less than an adaptive threshold were retained while the rest were zeroed out for each frame. The enhanced signal is obtained after applying the binary bling reverberant mask, estimated based on comparing features and threshold levels to the reverberant signal. Although the intelligibility is improved significantly there are some mistakes made by the binary reverberant mask makes mistakes in low frequency regions. These strategies were able to extract the direct sound from reverberated speech by discarding reverberated components but not able to suppress reverberant energies from late reflections.

3.6 Effect of background noise on speech perception with cochlear implants

Understanding speech in the presence of background interference and competing voices is a difficult task. It further complicates the process for individuals with sensorineural hearing loss. For hearing impaired and cochlear implant listeners, the speech perception in noise has continued to be challenging [3, 9, 20, 24, 25, 42, 83, 112]. The ability to successfully extract the information from the noisy speech depends on mechanism to identify the speech cues such as fundamental frequency (F0) and harmonic information. Normal hearing listeners can perform very well in noisy conditions as they are able to extract information from the modulating maskers. Although many studies on ideal binary masking demonstrated improved speech intelligibility in noise, access to clean and noise signals was assumed [41]. It was also noted that substantial computation resources that involve machine learning techniques were required to employ the ideal binary mask in the absence of information from clean and noise signals. [53] Without this the noise estimation techniques were shown ineffective in estimating binary masks. However there are issues such as weakly conveyed F0 cues, vowel and consonant boundaries, distortions in harmonic structure and sound localization for individuals with hearing loss. Two types of noise maskers are used to study the effect of additive noise on speech perception: speech-shaped noise and multi-talker babble noise.

3.6.1 Effect of noise maskers on speech perception

3.6.1.1 Speech-shaped noise

Speech-shaped noise has a long term average spectrum similar to that of a speech and is primarily used as a noise masker to study speech perception [79, 100, 102]. The advantage of using speech-shaped noise is that the SNR is held constant across the frequency range of speech [102]. Effect of speech-shaped noise on consonants and vowel recognition in cochlear implant listeners was

studies by Fu *et al.* [25]. To evaluate the performance in the presence of noise, they corrupted the vowel and consonant stimuli from Hillenbrand *et al.* [36] using speech-shaped noise at various signal-to-noise ratios between 24 and -15 dB SNR. The mean vowel recognition scores dropped from 66% in the quiet listening condition to 27% at 0 dB SNR. The mean consonant recognition scores dropped from 70% in quiet listening condition to 37% at 0 dB SNR. In another study, Friesen *et al.* [24] investigated the speech recognition as a function of spectral channels when speech-shaped noise is added to the stimuli. The stimuli consisted of sentences from HINT database [79] containing lists of ten sentences. They conducted experiments for 15, 10, 5 and 0 dB SNR levels, in a sound-proof room with the test material presented over a loud speaker via a compact disk player. The sentence recognition dropped uniformly from 85% in clean listening condition to 60% at 10 dB to 40% at 5 dB to 10% at 0 dB SNR levels respectively. They also noticed that, as the number of channels/electrodes was increased above seven to eight, word and sentence recognition continued to increase in normal-hearing listeners.

3.6.1.2 Multi-talker babble noise

While the speech-shaped noise has been used traditionally as a competing noise source, it does not always represent the type of noise encountered by individuals in everyday listening environments. The use of competing talkers (multi-talker babble) is another effective masker, and a realistic representation of the type of competing noise commonly encountered in everyday listening environments. A study by Fetterman and Domico [20], reported the extent of degradation in speech recognition in presence of multi-talker babble noise. They studied cochlear implant users using CIS, SAS and SPEAK signal processing strategies. The test stimuli was corrupted noise using multi-talker babble noise at 10 dB and 5 dB levels of SNR. All the experiments were conducted in an acoustically enclosed chamber and stimuli presented to the cochlear implant. The order of

presentation of the test material was randomized across the three different test conditions. The mean sentence recognition scores in quiet listening condition were 82.1% and significantly dropped to 73.04% in the presence of multi-talker babble noise at 10 dB SNR and to 47.36% at 5 dB SNR.

3.7 Noise reduction methods implemented for cochlear implants

Several noise reduction methods have been proposed over the years to enhance the quality of speech in the presence of background noise [38]. Most of these algorithms were based on either some kind of assumption or preprocessing. Berouti *et al.* [5] proposed a spectral noise subtraction method that is based on over subtraction to reduce musical noise. The signal is enhanced by performing the inverse Fourier transform of the square root of the obtained power spectrum after spectral subtraction combined with the phase of the noisy signal. Another technique for speech enhancement is Wiener filtering where a minimum mean square error (MMSE) estimator is determined from the noisy speech. MMSE wiener filter is used to obtain the enhanced speech signal from the noisy signal. Ephraim and Malah [18] used a minimum mean square error estimation of the spectral amplitude to reduce noise, since the value of the *a priori* SNR is not available, as it based on the clean speech signal spectral amplitude. They extended the method for the estimation of the a priori SNR. It was a decision directed approach to compute a posteriori SNR from noisy signal spectral amplitude and the noise spectral estimate. Extensive research has been done in the general area of speech enhancement over the last three decades, but not as many in the area of cochlear implants. In this section we discuss some of these speech enhancement techniques relevant to the techniques implemented for cochlear implants.

3.7.1 Adaptive beam filtering

Some of the early research to perform noise reduction for cochlear implants was done using adaptive beam forming that requires two microphones. Hamacher *et al.* [32] used two-channel adaptive beam forming techniques for performing noise reduction for cochlear implants. Four cochlear implant recipients who were users of Cochlear Corporation's Spectra processor participated in these studies. They reported that using the beam forming techniques the SNR was improved by about 6 dB. Van Hossel and Clark [106] also used the adaptive beam forming to perform noise reduction for cochlear implants. Four cochlear implant patients who were using spectral maxima signal processing strategy participated in these experiments. The test material consisted of sentences corrupted by multi-talker babble noise at 0 dB SNR. The adaptive beam forming was performed using two microphones. The input signals from the two microphones were added and subtracted to create the 'sum' and 'difference' signals. The 'difference' signal, which corresponds to the noise, was minimized using the least mean square (LMS) error criterion. The adaptive beam forming was implemented using an adaptive finite impulse response filter whose coefficients were updated using the LMS criterion. The four subjects were tested on the adaptive beam forming strategy and a reference strategy. The reference strategy was implemented by simply adding the two microphone input signals.

All the experiments were conducted in a sound proof chamber. The target sentence material was presented by a loudspeaker directly in front of the cochlear implant subject and multi-talker babble noise was presented at 90^{o} to the left of the subject. The subjects were tested on a total of four conditions that included the adaptive beam forming strategy and the reference strategy both in quiet and 0 dB SNR conditions. The subjects were tested on a list of fifteen sentences on each condition. The block of four experiments was repeated three times with one week time gap between each block of test. The mean sentence recognition was about 80% in quiet listening

conditions using both the strategies. Addition of multi-talker babble noise at 0 dB SNR resulted in mean sentence recognition of 10% and 40% in the case of the reference strategy and the adaptive beam filtering strategy respectively. Thus the noise reduction performed by the use of adaptive beam filtering resulted in a gain of 30% in sentence recognition over the reference condition.

3.7.2 Nonlinear spectral subtraction approach

This study was developed by modifying some of features of spectral subtraction method developed for normal hearing individuals. Yang and Fu [111] used nonlinear approach for improving speech perception in presence of background noise for cochlear implant users. In this method, input speech frame is divided into sub-blocks in order to reduce the variance. If the signal corrupted by noise $y\ (n)$ can be represented as sum of speech signal $s(n)$ and noise signal $d\ (n)$, then,

$$y\ (n) = s(n) + d(n) \tag{3.22}$$

The spectral subtraction operation corresponds to a time varying filtering operation was represented in terms of estimated spectral magnitude of short-term speech, noise and noisy speech $\hat{S}_N(f,i)$, $\widehat{D}_N(f,i)$ and $\hat{Y}_N(f,i)$, respectively.

$$\left|\hat{S}_N(f,i)\right| = G_N(f,i)\left|\hat{Y}_N(f,i)\right| \tag{3.23}$$

The gain function $G_N(f,i)$ can be represented as,

$$G_N(f,i) = \left(1 - k\frac{|\bar{D}_N^a(f,i)|}{|\hat{Y}_N^a(f,i)|}\right)^{1/a} \tag{3.24}$$

Where k is the subtraction factor and a determines the sharpness of the transition from when $G_N(f,i) = 1$ the spectral component is not modified and when $G_N(f,i) = 0$ when the spectral component is suppressed. The over subtraction factor is estimated as follows:

$$k(f,i) = \left(\frac{\frac{\max_{i-20 \le \tau \le i}(|D_M(f,i)|)}{|D_M(f,i)|}}{1+\gamma\left(\frac{|\hat{Y}_M(f,i)|}{|D_M(f,i)|}\right)} \right) \tag{3.25}$$

The variations may be decreased by using an adaptive exponential averaging of the gain function, where $\alpha_1(i)$ is an adaptive averaging time parameter derived from spectral discrepancy measure and $G_{M,1}(f,i)$,

$$\bar{G}_{M,1}(f,i) = \alpha_1(i)\,\mathrm{X}\,\bar{G}_{M,1}(f,i-1) + (1-\alpha_1(i))\,\mathrm{X}\,G_M(f,i) \tag{3.26}$$

$$\alpha_1(i) = \beta(i) \tag{3.27}$$

The spectral discrepancy measure, $\beta(i)$, depends on the relation between the current block spectrum, $\hat{Y}_M(f,i)$, and the current averaged noise spectrum, $\bar{D}_M(f,i)$,

For the gain function to adapt to the stationary input, the averaging time constant was computed as,

$$\beta(i) = \min\left\{\frac{\sum_{f=0}^{M-1}\left||\hat{Y}_M(f,i)| - |\bar{D}_M(f,i)|\right|}{\sum_{f=0}^{M-1}|\bar{D}_M(f,i)|}, 1\right\} \tag{3.28}$$

$$\bar{G}_{M,2}(f,i) = \alpha_2(i)\,\mathrm{X}\,\bar{G}_{M,2}(f,i-1) + (1-\alpha_2(i))\,\mathrm{X}\,G_M(f,i) \tag{3.29}$$

The variance reduced gain function is further subjected to smoothing as given as,

$$\bar{G}_{M,3}(f,i) = \max(\beta_{floor}\,\bar{G}_{M,2}(f,i)) \tag{3.30}$$

To reduce the musical noise in the signal, spectral flooring was used as,

$$\alpha_2(i) = \begin{cases} \gamma_c\,\alpha_2(i-1) + (1-\gamma_c)\alpha_1(i), & \alpha_2(i-1) \le \alpha_1(i) \\ \alpha_1(i)\text{ , Otherwise} \end{cases} \tag{3.31}$$

In this method the following values were used for the various constants in the given order k_c = 1.8, γ_c = 0.8, γ = 0.3, β_{floor}= 0.1. Finally, the enhanced spectrum is obtained using interpolation as given by the following equation:

$$\left|\hat{S}_N(f,i)\right| = G_{M\uparrow N,3}(f,i)\left|\hat{Y}_{L\uparrow N}(f,i)\right| \tag{3.32}$$

The enhanced signal in time domain is obtained by combining the enhanced spectrum with the noisy phase followed by inverse FFT.

Performance evaluation of this noise reduction method was done using seven cochlear users. Test stimuli was selected using HINT database [79] and processed with and without the use of the noise reduction algorithm using speech-shaped noise and multi-talker babble noise at 9, 6, 3 and 0 dB SNR levels. For speech-shaped noise, mean percent sentence recognition over all subjects and noise levels using the noise reduction algorithm was significantly higher by about 20% compared to unprocessed stimuli. For multi-talker babble noise, mean percent sentence recognition over all subjects and noise levels was not significantly greater (5% - 8%) than compared to unprocessed stimuli. Although the algorithm produced better results for speech-shaped noise, a better optimization of time constant values may be helpful in competing speech environments.

3.7.3 Subspace approach

Loizou *et al.* [67] proposed a noise reduction algorithm based on subspace technique for cochlear implant users. An extended subspace approach for colored speech-shaped noise was proposed [38]. Subspace approach relies on the basic principle of decomposition of the corrupted speech vector into ‘signal’ subspace and ‘noise’ subspace respectively. In the subspace approach, clean signal is estimated by removing the signal components from noise subspace and by retaining the signal

components in signal subspace. In their approach, *y* is considered as the noisy vector and $\hat{x}$ is the estimate of clean signal vector respectively.

$$\hat{x} = H \,.\, y \tag{3.33}$$

Where H is the transformation matrix. When this matrix is applied to the noisy signals yields the estimate of the clean. The estimation error can be represented as follows:

$$\varepsilon = \hat{x} - x = H \,.\, y - x \;=\; (H - I).x + H.n \tag{3.34}$$

Where n is the noise vector. Since the transformation matrix H is not perfect, it introduces speech distortions represented by the term $(H - I).x$ and noise distortion represented by the term $H.n$. As the speech and noise distortion from (3.34) are decoupled, the optimal transformation matrix H that would minimize the speech distortion can be developed subject to a preset threshold for the amount of noise distortion. The solution to this constrained minimization problem was determined for colored noise [38]:

$$H = V^{-T}\Lambda(\Lambda + \mu\mathrm{I})^{-1}\,V^{-T} \tag{3.35}$$

In the above equation, μ is a Lagrange multiplier (typical value between 1 -20) used in constrained minimization problems. V is an eigenvector matrix that decomposes the corrupt signal into signal and noise subspaces. Λ is a diagonal eigenvalue vector obtained from noisy speech vector. $\Lambda(\Lambda + \mu\mathrm{I})^{-1}$ is a diagonal matrix that multiplies the signal component with the gain and zero out the noise component. Finally, V^{-T} performs the inverse transform of the signal. They implemented the above signal subspace algorithm by computing the estimate of the clean signal vector ($\hat{x} = H_{opt} \,.\, y$) by transforming the noisy speech frame based on (3.35) and then providing the estimated signals $\hat{x}$ as input to the CI processor.

The test stimuli was choosen to be HINT sentences [79] corrupted with 5 dB speech-shaped noise. Fourteen CI subjects were tested for sentence recognition and the mean percent correct

scores obtained with unprocessed and preprocessed sentences by the subspace algorithm were 19% and 44%. Some subjects showed about 50% improvement with the proposed approach. Although the results obtained were significant, the above subspace algorithm was only tested in speech-shaped (stationary) noise. The performance with multi-talker babble noise or nonstationary environments is not evaluated. Perhaps a better noise estimation algorithm would help with the stationary and nonstationary environment.

3.7.4 Conclusion

Several noise reduction algorithms have been discussed that suppress noise to enhance the intelligibility. Most of the above algorithms tend to improve the speech quality; however, they are preprocessing based and require optimization to reduce noise components. For instance, adaptive beam filtering approach present a higher capability of interference cancellation, but they are much more sensitive to steering errors and suffer from signal leakage and degradation. On the other hand, spectral subtraction based noise reduction technique worked well for stationary noise masker than the speech babble. In some cases, the spectral subtraction might produce speech like time-varying characteristics and further decrease intelligibility. Another enhanced algorithm that addressed the concerns in spectral subtraction is the subspace approach. Although this method eliminates the musical noise in reconstructed speech, it also applies to stationary noise and better noise estimation algorithm is required to improve the performance. With many other noise reduction methods, the process of filtering and suppressing noise results in degraded harmonic structure which results as unwanted distortions such as musical noise. Therefore, a mechanism to restore the harmonic structure might help regain the intelligible information in the signal after the noise is suppressed.

CHAPTER 4

PERCEPTION OF TELEPHONE-PROCESSED SPEECH WITH COMBINED ELECTRIC AND ACOUSTIC STIMULATION

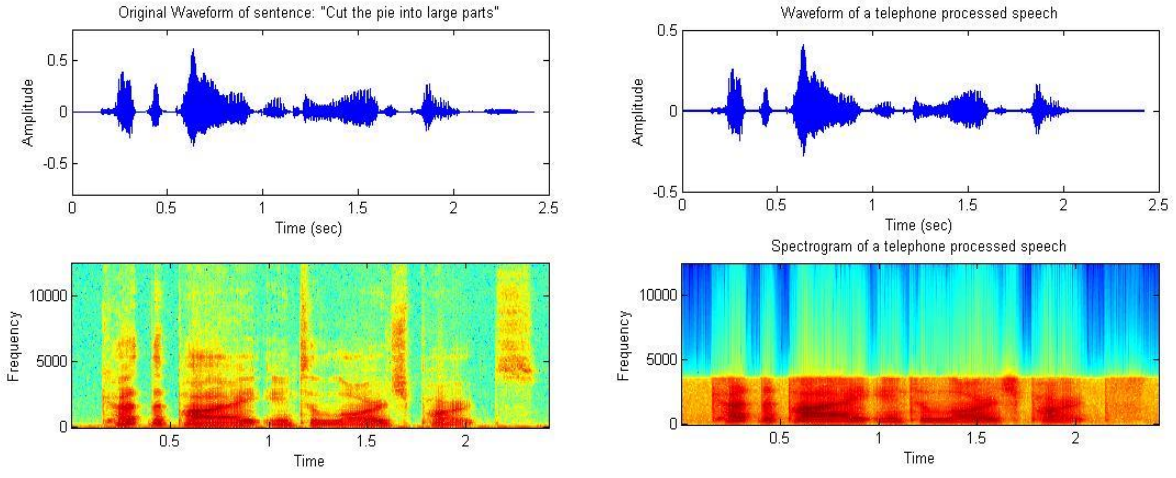
For hearing impaired listeners, communication via telephone may be convoluted due the reduced audibility of the telephone signal even in quiet environments. In this chapter, we will develop a strategy to assess the individual effects of adding low- or high-frequency information to the frequency-limited telephone processed speech on perception of cochlear implant and bimodal listeners in quiet listening condition. We propose presenting the cochlear implant and bimodal listeners with four different configurations of speech: wideband speech (WB), bandpass-filtered speech (frequencies between 300 Hz and 3400 Hz, BP), high-pass filtered speech (frequencies above 300 Hz), and low-pass filtered speech (frequencies less than 3400 Hz,) in quiet listening condition.

4.1 Introduction

In order to make telephone calls, we utilize the public switched telephone network (PSTN) which transmits frequencies between 300 Hz to 3400 Hz. Based on a study on normal hearing subjects, speech intelligibility tests revealed an effective frequency bandwidth, which was selected for telephone speech [22]. This frequency limitation was implemented to reduce the cost and save bandwidth of communication channel by the public switched telephone network (PSTN). Due to the reduced frequency spectrum, frequency information below 300 Hz and above 3400 Hz is severely distorted in telephone processed speech. The band-limiting presents trivial impact on normal hearing listeners; however, for hearing impaired listeners, this reduced spectral information significantly degrades the speech intelligibility. In this chapter, we will investigate the effects of

adding or restoring low- or high- frequency information in the band-limited telephone-processed speech for cochlear implant and bimodal listeners and evaluate the performance in quite listening condition. In the proposed study, bimodal users were presented with wide-band speech (WB), bandpass-filtered (300-3400 Hz) telephone speech (BP), high-pass filtered (f>300 Hz, HP) speech (i.e., distorted frequency components above 3400 Hz in telephone speech were restored) and low-pass filtered (f<3400 Hz, LP) speech (i.e., distorted frequency components below 300 Hz in telephone speech were restored) under quiet listening condition to study their perceptual effects.

From prior research studies, it is evident that use of telephone for conversations is difficult for CI listeners and that the telephone speech recognition is significantly worse compared to their recognition of wideband speech. Results from all these studies has led to the generalized understanding that the reduced bandwidth of the telephone speech accounts for a significant amount of performance deterioration [12, 12, 26, 37, 47, 71]. As discussed in the literature review, most of these studies provide support for the idea of developing bandwidth extension techniques that for telephone speech which is expected to enhance the CI listeners' telephone usage. Figure



(a) (b)

Figure 4.1. Effects of frequency-limiting in telephone processed speech. (a) Wideband speech spectrum, (b) T**e**lephone processed (band-limited) speech spectrum.

4.1 shows the effect of band-limiting on clean speech. As it can be seen that the frequency components below 300 Hz and above 3400Hz are distorted.

To improve the perception of telephone speech by CI listeners, few researchers had investigated techniques to extend the bandwidth of the telephone networks to study its impact on CI listeners. Existing studies were able to examine the performance and perception of broadband and band-limited (frequency-limited) speech among CI listeners [26, 37]. However, these studies were unable to address the individual effects derived from low frequencies below 300 Hz and that from high frequencies above 3400 Hz to study their perceptual effects. Since current CI coding strategies based on envelop-vocoding processing provide weakly conveyed fundamental frequency (F0) cues, low frequency acoustic information (i.e., f<300 Hz) might not be accessible and therefore, it makes sense to follow the practice. However, severely distorted frequency components can significantly impact the formant and harmonic representation of vocoded telephone speech. Another study evaluated the telephone speech perception in Mandarin speaking CI listeners [37], the effect of telephone speech was highly variable among subjects and the results suggested that the variability was due to the lack of high frequency information in telephone speech. The contributions from low-frequency speech cues were not evaluated in their study.

This chapter is organized in the following order. In section 4.2, the benefits of combined electric and acoustic stimulations for telephone speech perception is discussed. The proposed model and signal processing involved for this study is presented in section 4.3. Section 4.4 describes the implementation details of subjective listening tests, stimuli used and procedure followed in this study. Analysis and experimental results are given in Section 4.5. Summary and conclusions are presented in Section 4.6.

4.2 Combined electric and acoustic stimulation for band-limited speech perception

Combined electric and acoustic stimulations have recently gained interest in the field of cochlear implants and many researchers have taken advantage of EAS to study the contributing factors of hearing aids to benefit the CI listeners. Hearing impaired listeners with low-frequency residual hearing can be fitted with cochlear implant in one ear and hearing aids in the contralateral ear (called Bimodal fitting) or in the same ear using a short electrode (called Hybrid CI). As discussed in Chapter 2, many studies have documented the improved speech intelligibility with combined electric and acoustic stimulations, especially in noise. When low-frequency acoustic sound from hearing aid is added to the mid- to high- frequency electrical information presented by the CI, it provides added benefits from both the devices. Better representation of F0 cues from the low-frequency acoustic energy, improved glimpsing and harmonic representation accounts for the improved intelligibility [9, 42, 103, 114]. Fundamental frequency (F0) cues and harmonics structure has been shown to be a useful in identifying word boundaries, vowel transitions that facilitate lexical segmentation and tone identification for cochlear listeners [64, 91].

The rationale in this chapter has three fold: (1) by assessing and differentiating the effects of added low- and high- frequency information, the present study provides support for development of efficient bandwidth extension techniques; (2) with improved representation of F0 cues, bimodal listeners may benefit from adding low-frequency information (i.e., f<300 Hz) to band-limited telephone speech, and (3) since hearing aids could provide low- frequency acoustic information, it helpful to examine whether hearing aids benefits cochlear implant listeners in recognizing band-limited telephone speech. This could possibly allow for improved perception in the frequency range between 300 Hz and 600 Hz, due to the low frequency information overlapping with the bandwidth of telephone speech.

From prior studies, there is enough support that bandwidth extension (i.e., low- or high-frequency extension) would yield performance benefits [47, 63, 81]. Experiments with bimodal listeners is expected to enable the present study to address the question of whether the efforts should be focused on extending the frequencies below 300 Hz or the frequencies above 3400 Hz.

4.3 Proposed approach for bandwidth extension

4.3.1 Overview of the proposed model

Figure 4.2 presents the model designed for this proposed study. Four different filtering configurations are applied to test the CI listening. LP refers to the telephone processed speech with distorted low frequency components restored, HP refers to the telephone processed speech with distorted high frequency components restored, BP refers to the telephone processed speech by itself and WB refers to the clean signal.

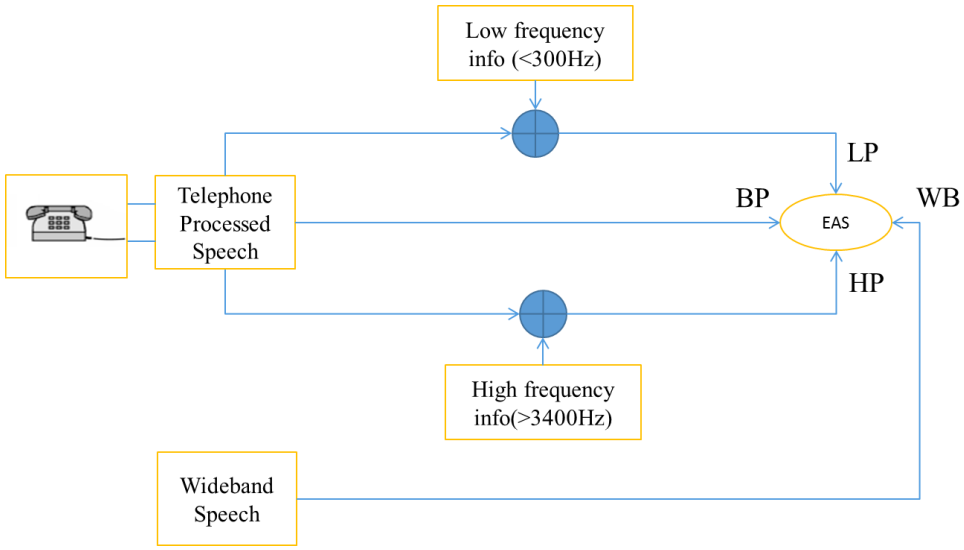


Figure 4.2. Block diagram of the proposed model to assess the contributions of low and high frequency information.

Contribution from low frequencies below 300 Hz are studied from LP condition, and those from high frequencies above 3400 Hz are studied from HP condition. By individually assessing these two conditions, this study expects to gain a better understanding of whether low frequency

information or high frequency information is beneficial to cochlear implant and bimodal listeners. Assess the contributions of the low- and/or high-frequency information to telephone speech could provide support and direction for future efforts in developing efficient bandwidth extension techniques that could concentrate on extending frequency band towards either low or high frequencies for improved telephone speech perception by cochlear implant listeners

4.3.2 Signal processing

Modified Intermediate Reference System (IRS) filter [48] is used to filter the original wideband speech signals into three different configurations. The frequency response of the IRS filter is discussed in Appendix A. There are many different applications of these IRS filter in telecommunication applications, one of which is to implement noise reduction [38]. The IRS filter is used to simulate the frequency characteristics of the received signals in telephone handsets. Therefore, we use the filter to generate four filtering configurations. Magnitude gain values were designed to simulate the response of IRS filter, such that the gain increases from -173.6 dB to -1 dB (-173.6, -52, -32, -24, -18, -12, -8, -6, -4, -2, -1) for frequencies between 0 Hz to 488 Hz (0, 38, 88, 113, 148, 188, 238, 288, 338, 388, 488) and reaches maximum at 0 dB for frequencies between 588 Hz and 3238 Hz and then begins to drop to -8 dB at 3488 Hz and continue to -212 dB at 3988 Hz, respectively. A simulated low-pass filter is used to examine the contributions of low-frequency information (below 300 Hz) and a simulated high-pass filter is used to examine the contributions of high-frequency information (above 3400 Hz). The frequency response of band-pass, low-pass and high-pass filters are shown in Figure 4.3. The upper panel show the frequency response of the band-limited modified IRS filters used by ITU [48]. The middle panel depicts the frequency response of the low-pass filtering network, which has frequency components from 300 Hz to 3400 Hz and retained (no distortion) lower frequencies below 300 Hz; the bottom panel

shows the frequency response of the high-pass filtering network, which has frequency components from 300 Hz to 3400 Hz and retained (no distortion) higher frequencies above 3400 Hz. For the wideband speech, there is no distortion in any frequencies, therefore it uses an all-pass filter. Figure 4.4 shows the waveforms and spectrograms of all four filtering conditions.

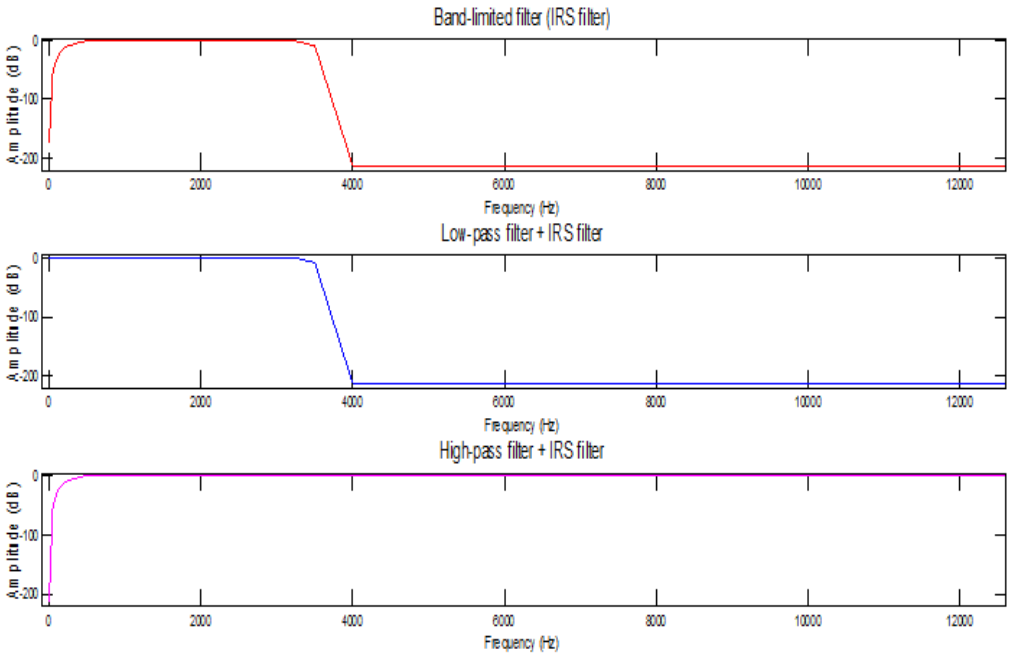


Figure 4.3. Frequency response of band-limited filter (top), low-pass filter (middle) and high-pass filter (bottom) simulated using IRS filter.

4.4 Implementation details

4.4.1 Subjective listening tests

Eleven post-lingually deafened subjects fitted with cochlear implant and hearing aids in the contralateral ear participated in the study. Table 4.1 provides the demographics of the subjects and from the table, it can be seen that all subjects have had at least one and a half years of experience with cochlear implant and 6 years of hearing aid experience at the time of testing. Also note, that

the average age of the subjects was about 61 years. The speech stimuli used in this study were sentences from IEEE database [49] obtained from Loizou [68]. There are 72 sentence lists in the IEEE database with 10 sentences each, which are phonetically balanced and recorded in a double-walled sound-attenuation booth at a sampling rate of 25 kHz. These sentences are produced by the male speaker. The mean F0 for these sentences is about 128 Hz, and the standard deviation is around 21 Hz.

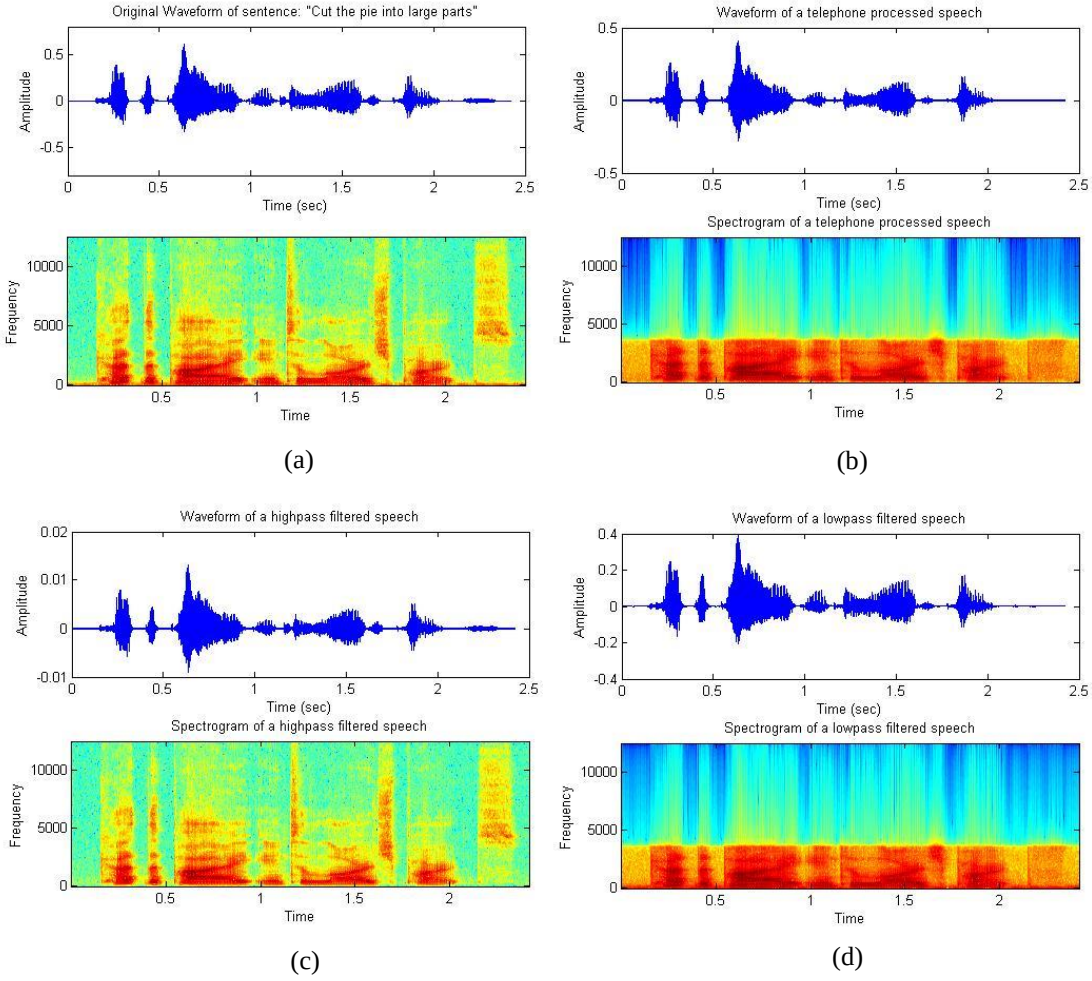


Figure 4.4. Waveforms and spectrograms of a speech sentence from IEEE database: "Cut the pie into large parts". a) Wideband speech, b) Band-limited speech, c) High-pass filtered speech (with no distortions in higher frequencies above 3400 Hz), d) Low-pass filtered speech (with no distortions in low frequencies below 300 Hz).

Subject	Age at testing (Years)	Gender	Etiology	CI Device	Implanted ear	Duration of hearing loss (NIE, Years)	Duration of hearing loss (IE, Years)	CI experience (Years)	HA experience (Years)
S1	51	F	Autoimmune disease of inner ear	Nucleus Freedom	R	7	11	1.5	6
S2	56	F	Unknown	MED-EL Sonata	L	13	3	2	8
S3	54	M	Genetic nerve deafness	AB Harmony	R	30	40	4	44
S4	67	M	Familial	Nucleus Freedom	R	61	61	5	55
S5	61	F	Unknown	AB Harmony	L	15	15	2	15
S6	67	M	Unknown	Nucleus Freedom	L	58	58	3	33
S7	69	M	Unknown	Nucleus Freedom	R	7	7	3	7
S8	82	M	Mixed Condition/SN Hearing	Nucleus Freedom	R	17	17	3	6
S9	60	M	Unknown	Nucleus Freedom	R	26	26	2.5	25
S10	33	F	Meniere's	Nucleus Freedom	L	14	16	2	6
S11	69	M	Noise exposure	Nucleus Freedom	R	20	20	1.5	26
Mean	60.8					24.4	25	2.7	21
Standard Deviation	12.6					18.7	19.7	1.1	17.2

Table 4.1. Demographic details of the bimodal listeners participated in the study.

4.4.2 Testing Procedure

The subjective listening tests were conducted in a double-walled sound-treated booth in quiet listening conditions. There were three listening modes and four filtering configurations designed for this study.

For listening mode (A), the subject used only hearing aid and the implant was removed. For listening mode (E), the subjects used only cochlear implant and hearing aid was removed. In order to exclude the residual acoustic hearing, an ear foam plug and a circumaural headphone (DAP World, Inc., Sherman Oaks, CA, USA) were used. Finally, for the listening mode (A+E), the subject used both cochlear implant and a hearing aid.The four filtering configurations were: WB (wideband speech), BP (band-limited or frequency-limited telephone speech), LP (low-pass filtered; lower frequencies below 300 Hz restored in telephone speech) and HP (high-pass filtered; higher frequencies above 3400 Hz restored in telephone speech). Therefore, a total of 12 (3 listening modes * 4 filtering configurations) testing conditions were designed. Two IEEE sentence lists were used per condition, therefore a total of 24 lists were used to test each subject. The test conditions were randomized for each subject and each sentence was presented once at a 65 dB SPL via loudspeaker placed at 1m in front of the listener. Speech recognition tests were conducted for all the three listening modes: A condition, E condition and A+E condition. The subjects were asked to leave the settings on hearing aids and implant the same as their daily settings.

The subjects participated in 30-minute practice sessions prior to the testing to help subject familiarize with the testing procedure and listening conditions. Different database was chosen to deliver the sentences for practice session, AzBio (Arizona State University, Tempe, AZ, USA) database [90]. The experimenter was present with the subject in the double-walled sound- treated

booth, and played the stimuli to the subject. The subjects were asked to speak verbally what they had heard, and the experimenter would record the response. The experimenter was allowed to replay the stimuli but not to provide any feedback to the subject during the testing.

4.5 Experimental results

Percentage of words identified correctly were computed and that would be the measure of the performance. (Articles such as a/an/the were not scored). Figure 4.5 shows the percent correct scores for the 11 subjects using hearing aid (A), cochlear implant (E) or combined hearing aid and cochlear implant (A+E). WB indicates the unprocessed (wideband) stimuli, BP indicates the band-pass filtered stimuli (i.e., 300 Hz<f<3400 Hz), LP indicates the low-pass filtered stimuli (i.e., f<3400 Hz), and HP indicates the high-pass filtered stimuli (i.e., f>300 Hz). The mean percent correct scores are shown in the Table 4.2. For the three listening modes (HA, CI and HA+CI) and four filtering conditions (WB, BP, LP, HP) the results are displayed in the Table 4.2. The results indicated significant variability across subjects in the HA and CI conditions. The scores shown in Figure 4.5 were first converted to rational arcsine units (RAU) using the rationalized arcsine transform [94] for the purpose of comparison across conditions.

Filtering	HA	CI	HA+CI
WB	29.20%	69.96%	79.58%
BP(Telephone Speech)	23.44%	53.75%	67.51%
LP	21.20%	58.99%	73.05%
HP	28.63%	67.37%	76.77%

Table 4.2. Mean percent correct scores for four filtering conditions across three listening conditions: WB, BP, LP and HP.

Benefits of acoustic stimulations from hearing aids (A) when combined with electric stimulations from cochlear implant (E) were evaluated using paired-sample t -tests between converted scores obtained with cochlear implant (E) and converted scores obtained with combined hearing aid and cochlear implant (A+E) with significance level of $\alpha = 0.05$. The results from the comparison test showed that listening to stimuli with both hearing aid and cochlear implant (A+E) was significantly better (with $p < 0.05$) than listening with cochlear implant alone (E) for all four types of filtered stimuli (WB, BP, LP and HP). The mean benefit of 12% was observed across all four filtering conditions.

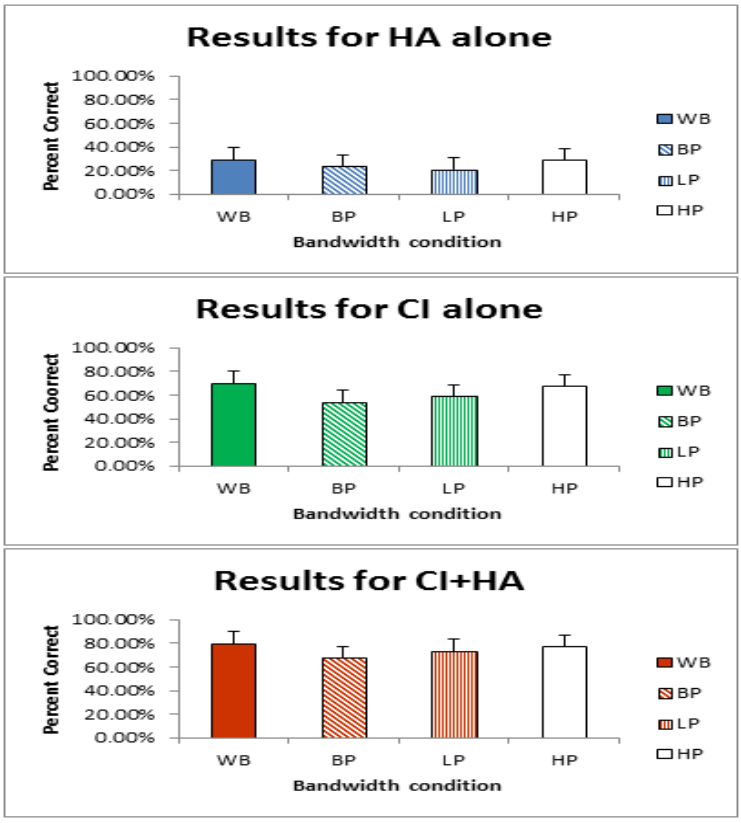


Figure 4.5. Percent correct scores from eleven subject for three listening modes: hearing aid only (A), cochlear implant only (E), and combined hearing aid and cochlear implant (A+E).

In order to assess the bandwidth effect on telephone speech perception by cochlear implant, multiple paired comparisons were performed with Bonferroni correction between converted scores

from all four filtering conditions for ‘E’ listening mode. The statistical significance level with Bonferroni corrections was set to $p < 0.0125$ with significance level of $\alpha = 0.05$. The results indicated no statistically significant differences between the LP and BP scores with $p > 0.1$, and between the WB and HP scores with $p > 0.2$. Although the WB scores were not significantly better than the LP scores ($p = 0.035$), HP scores proved to be significantly better than the BP scores with $p < 0.01$. This shows that restoring higher frequencies to telephone speech yields better speech recognition compared to band-limited telephone speech for cochlear implant listeners. Now, in order to assess the effect of bandwidth on telephone speech perception by hearing aid, multi-paired comparisons were performed again with Bonferroni correction between converted scores from all four filtering conditions for ‘A’ listening mode. The statistical significance level with Bonferroni corrections was set similar to the previous values ($\alpha = 0.05$). The results indicated no statistically significant differences between the LP and BP scores with $p > 0.2$, between the WB and HP scores with $p > 0.7$ and between the HB and BP scores with $p > 0.1$. Similar to results from ‘E’ condition, the WB scores were not significantly better than the LP scores ($p = 0.023$). Finally, to assess the effect of bandwidth on telephone speech perception by combined cochlear implant and hearing aid (A+E), similar multi-paired test were run as before but this time with Bonferroni correction between converted scores from all four filtering conditions for ‘A+E’ listening mode. The comparisons showed no statistically significant differences between the LP and BP scores with $p > 0.1$, between the WB and HP scores with $p > 0.1$. This time, the WB condition showed significantly better scores than the LP condition ($p = 0.001$). As in ‘E’ listening mode, HP scores were significantly better than BP scores ($p < 0.01$).

4.6 Summary and conclusions

In this chapter, the effect of bandwidth on speech perception with electric and acoustic stimulation was examined. This study helped in assessing whether low frequency and high frequency information play different roles for bandwidth extension to improve telephone speech perception by cochlear implant and bimodal listeners. This study also helps address whether hearing aids help facilitate perception of band-limited speech when combined with a cochlear implant. Some of the prior studies have shown that F0 was the beneficial cue for the bimodal benefits for speech recognition in noise by assessing the effect of low-passed acoustic stimuli [114] and high-passed electric stimuli [113]. The difference between those studies and the present study is that the bimodal users in those studies had access to high-frequency information.

From section 4.5, experimental results showed that hearing aid benefited speech perception in various band-limited conditions for bimodal users under quiet conditions (See Table 4.2). For telephone speech, which is the BP condition, bimodal (A+E) was significantly better than implant alone (E) by 13.76%. For low-pass filtered speech, which is the LP condition, bimodal (A+E) was significantly better than implant alone (E) by 14.06%. For high-pass filtered speech, which is the HP condition, bimodal (A+E) was significantly better than implant alone (E) by 9.4%. And, for wideband speech, which is WB, bimodal (A+E) was significantly better than implant alone (E) by 9.62 %.

Therefore, the addition of a contralateral hearing aid benefited CI hearing. Highest benefit (14.06%) was observed when low frequency information was present, whereas least benefit (9.4%) was observed when high frequency information was present. These results seem to be in line with the fact that hearing aids mainly facilitates access to lower frequency information. Figure 4.5 shows poor aided thresholds in high frequencies for many subjects, however, speech recognition

for several subjects for hearing aid (A) condition was better for high frequency information. Addition of a contralateral hearing aid showed significant benefit 13.76% to CI listening for telephone speech (BP condition). This benefit could be attributed to an increased spectral resolution (may be an overlap between hearing aid and cochlear implant hearing), therefore, further investigation into noisy conditions could help address questions whether bimodal listeners could take advantage of the information provided by the hearing aids with mostly missing distorted F0 information and whether the results stay similar to the clean stimuli. If the intelligibility scores for noisy stimuli show the same pattern, then there could possibly be underlying mechanism in addition to F0 that might explain the bimodal benefits.

Consistent with earlier findings [37], results from this study have shown that lack of high frequency information reduces CI listeners' speech recognition. Therefore, by observing both conventional CI hearing and bimodal hearing scores, it is obvious that low and high frequency information play different roles in speech recognition under quiet listening conditions. In summary, adding low frequency information to telephone speech have least to no impact on CI and bimodal listening, whereas adding high frequency information to telephone speech significantly improved speech recognition for both CI and bimodal listening. Since these results are applicable to quite listening conditions, it is expected that low frequency information to telephone speech will benefit bimodal users in noisy conditions due to the better F0 representation and improved glimpsing facilitated by the use of a hearing aid [62 114]. The results of this study were partly published in [44] and provide support for the design of efficient bandwidth extension techniques that would extend higher frequency information, at least in quiet environments. However, for noisy listening condition, results might vary due to other underlying mechanisms and therefore, caution needs to be exercised to draw any conclusions.

In conclusion, the study presented in this chapter complements the prior ones that showed that the low frequency cues are not useful on their own but they are beneficial when provided to bimodal listeners via HA. Although simply extending the bandwidth of speech below 300 Hz might not be useful for both CI and bimodal listening, this study suggests that delivering this low-frequency information via a different mode could possibly improve speech recognition. Eventually, it is expected that improved coding strategies would consider extending the bandwidth of telephone speech toward both low and high frequencies and therefore, these findings provide guidance and support for the design of efficient bandwidth extension algorithms.

CHAPTER 5

EVALUATION OF SPECTRAL SUBTRACTION STRATEGY TO SUPPRESS REVERBERANT ENERGY IN COCHLEAR IMPLANT DEVICES

Acoustic reverberation is a common phenomenon observed in almost all enclosed spaces. Reverberation is known to have negative impact on speech intelligibility, especially for hearing impaired listeners. The reverberant sound in any enclosed space reaching the listener's ears consists of the direct sound, early reflections, and late reflections. Late reflections are considered to be detrimental to speech intelligibility by cochlear implant listeners. The smearing effect of room reverberation can significantly impair the ability of cochlear implant listeners to understand speech. To ameliorate the effects of reverberation, current dereverberation algorithms focus on retrieving the direct sound from the reverberated signal by inverse filtering the reverberation. In this chapter, we propose a strategy based on spectral subtraction that is capable of suppressing late reflections in speech perceived by CI listeners. We also, compare the spectral subtraction strategy against ideal reverberant (binary) masking approach.

5.1 Introduction

Acoustic reverberation is the sum of all sound reflections arriving at the listener in an acoustical enclosed space after the enclosure has been excited by an impulsive sound signal [58]. Additive reverberant energy can cause significant deterioration in speech intelligibility when microphones or human ears are placed at a relatively large distance from the target sound source. From the literature review in chapter 3, it is known that prior studies have suggested that reverberation flattens formant transitions in vowels, results in weak-energy speech units being masked by preceding segments with strong energies, smears spectral cues, reduces temporal amplitude

modulations, and increases low-frequency energy, which consequently masks high-frequency components [3, 6, 75, 77]. The theory of speech masking by acoustic reverberation was initially preoposed by Bolt and MacDonald in 1949 [7]. They computed the articulation index by assuming the speech signal as a series of discrete pulses of acoustic energy separated by relatively quiet gaps, and thereby measuring the intelligibility of reverberated speech by focusing mainly on overlap-masking effects. These effects were introduced by energy components of previous pulses masking the components of the following pulses. In this contribution, we propose a strategy to ameliorate the negative effects of reverberation on speech. They assumed that the majority of the intelligibility was carried in the leading edge of the pulse. It was clear that self-masking was mainly caused by the energy components of the leading edge of the pulse masking the remaining portions of the same pulse, which carried only minimal intelligibility in the first place. Therefore, it was suggested that self-masking effects (energy smearing within a phoneme) would have little to no effect on speech intelligibility. Their model yielded percent articulation values which agreed precisely with values determined experimentally at reverberation times of < 2s. In another study [77], the authors used /s/-spectrum-shaped noise to compare the identification of consonants following /s/ under reverberated conditions with the identification of the same consonants without the preceding /s/ phoneme under noise masking conditions. Since there was no self-masking effects, the error patterns caused by reverberation could be comparable to those caused by the noise masker. As this was not the case, the authors ultimately concluded that both overlap-masking and self-masking had a negative impact on consonant identification.

Poissant *et al.* [85] showed the percent correct recognition scores deteriorate significantly, when that when only a small number of spectral bands is used to vocoded speech signals inside a simulated reverberant field. In another study [108], reverberated speech vocoded using 6 channels

yield the lowest intelligibility scores when compared to speech recognition for reverberant stimuli vocoded using 12 or more channels. It was found that intelligibility was mildly affected by RT_{60} and the source-to-listener distance in the latter case. Sentence recognition was assessed with six adult listeners wearing CI devices using varying degrees of reverberation by Kokkinakis *et al.* [55]. Their study was the first to show that speech intelligibility by individuals who relied exclusively on CIs decreased exponentially with a linear increase in reverberation time. These findings were also later confirmed by results obtained from [33], where speech intelligibility by CI listeners in two moderate-to-severe reverberation conditions (RT_{60} = 0.6 s and 0.8 s) were tested. It was shown in both the studies that intelligibility of electrically processed sentences may drop by approximately 60% points or more on average when reverberation time increases from RT_{60} = 0 s to RT_{60} = 0.8 s

This chapter is organized as follows. Section 5.2 presents the motivation to target the late reverberation that is detrimental to speech intelligibility. In Section 5.3, the idea of reverberation suppression strategy is discussed. Section 5.4 presents the overview of the proposed discussing the statistical model of room impulse response and the algorithm formulation of spectral subtraction strategy. Section 5.5 explains the implementation of reverberation suppression based on SS with the help of procedure and speech stimuli used. Comparison of SS and Ideal Reverberant Masking (IRM) strategies is presented in Section 5.6 along with the experimental results and analysis. Summary and Conclusions are discussed in Section 5.7.

5.2 Motivation

Speech degradation due to reverberation has led many researchers to investigate dereverberation algorithms for speech enhancement. Dereverberation often relies on inverse filtering and, hence, seeks to “invert” the reverberation process. This was one of the first and remains one of the most

commonly used methods today. However, this technique relied on the prior knowledge of the room acoustic impulse response and was sensitive to the changes in the overall impulse response caused by even slight sound-source movements [56]. Therefore, from the literature it is clear that researchers investigated other processing strategies to suppress reverberation in CI devices [34, 35, 55]. An ideal reverberant masking (IRM) approach was developed which was based on the estimation of the signal-to-reverberant ratio (SRR) of the individual spectral channels. Shortly after, the ideal binary masking approach was extended into a blind one by computing binary masks without prior knowledge of the SRR [34]. All of these methods resulted in significant intelligibility improvement of reverberated speech perceived by CI listeners.

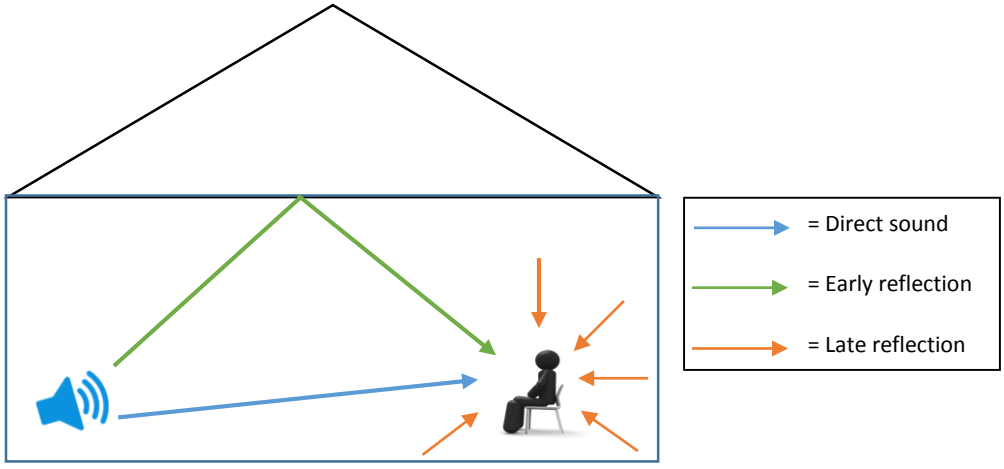


Figure 5.1. A block diagram of room acoustics and reverberated signals.

Most dereverberation algorithms focus on recovering the direct sound from the reverberated signal. This approach might be suitable for machine applications but for human listeners this has limitations. The reverberant sound in any enclosed space reaching the listener's ears consists of the direct sound, early reflections, and late reflections as seen in Figure 5.1. From literature in chapter 3, early reflections are considered beneficial to the overall speech

intelligibility, as they can be perceptually integrated with the direct sound and late reflections are considered to be detrimental to speech intelligibility. Since early and late reflections will likely produce different perceptual effects on the human auditory system [4, 30], reverberation suppression strategies need to be able to efficiently differentiate between these two types of reverberation. This idea drives the motivation of this study is to differentiate the two types of reverberations by retaining those that are beneficial while suppressing those that are detrimental to speech intelligibility.

5. 3 Reverberation strategy for suppressing late reverberation

Earlier findings provided strong evidence that reverberated speech caused by the late part of additive reverberation is much more detrimental to overall intelligibility than early reverberation [45]. This work is motivated by the two studies [34, 55]. In this chapter, we examine an efficient spectral-subtraction based strategy for reverberation suppression, which aims to remove the late reverberation components from the room impulse response (RIR) filter. Our rationale is twofold: (1) since late reflections account for the worsened speech intelligibility observed in CI listeners [45] we propose to target only reverberant energies caused by these late reflections; and (2) using spectral subtraction strategy which is an efficient alternative to binary T-F strategies for reducing the masking effects caused by additive reverberant energy. By modeling the overlap-masking effects caused by late reflections as an uncorrelated random noise process, we investigate whether spectral subtraction strategy could be beneficial for CI listeners. The major difference between this strategy and those that are developed earlier is that in this contribution we focus explicitly on the suppressing late reverberations by the use of spectral subtraction (SS).

Spectral subtraction strategy can operate directly on the reverberant signal envelopes, unlike binary time-frequency (T-F) masking strategies, which compare the energy of the

reverberant speech with that of the anechoic speech in order to determine whether to retain or discard a T-F unit. The central hypothesis is that speech intelligibility improves significantly by reducing the energy due to late reverberation components through SS processing. We hope that the SS strategy proposed in this work may allow for real-time implementation in clinical CI processors.

5.4 Algorithm Overview

A statistical model of the RIR filter was first developed by Polack [86]. The application of this model to machine speech recognition scored modest success [61]. This statistical model is described in the section below.

5.4.1 Statistical model for room acoustic impulse response

Statistical reverberation model developed by Polack [86] was used in discrete-time domain in this study. Since the acoustic behavior in real rooms is too complex to model explicitly, statistical room acoustics is often used. Polack developed a time-domain model complementing Schroeder's frequency domain model. In this model, an acoustic impulse response is described as one realization of a non-stationary stochastic process. This model is defined as:

$$h[n] = \begin{cases} b[n]\, e^{-\zeta(n)}, & n \geq 0 \\ 0 \quad , & \text{Otherwise} \end{cases} \tag{5.1}$$

where n denotes the discrete time index, $b[n]$ is a zero-mean stationary Gaussian noise sequence, and ζ is linked to the reverberation time RT_{60}. Reverberation time is frequency dependent due to the frequency dependent coefficient of walls and frequency dependent absorption coefficient of air. It should be noted that Polack's model is valid in cases for which the distance between the source and the measurement point is greater than the critical distance. In some cases where the source to receiver distance is larger than the critical distance, Polack's statistical model although

useful may not be accurate for the room impulse response. In the current study, we use this model in the discrete-time domain to investigate whether a reverberation suppression strategy based on SS can improve the perception of reverberated speech by CI listeners.

5.4.2 Spectral Subtraction strategy

Let *s*[n] denote the clean discrete-time speech signal, and *h*[n] denote the RIR filter. Then the reverberated speech signal, *x*[n], is obtained by:

$$x[n] = s[n] \otimes h[n] \tag{5.2}$$

Where $\otimes$ indicates the discrete-time convolution operator. The causal RIR filter *h*[n] can be decomposed into three separate components: *h*[0], which represents the direct path, h_e[n], which represents the early reflection, and h_l[n], which represents the late reflections. A simplified variation of the statistical model for the RIR filter in [86] can be described as a random process with an exponentially decayed envelope signal as:

$$h[n] = \begin{cases} 0, & \text{for } n < 0 \\ h[0], & \text{for } n = 0 \\ h_e[n], & \text{for } 1 \leq n \leq T_e f_s - 1 \\ h_l[n], & \text{for } T_e f_s \leq n \leq N - 1 \end{cases} \tag{5.3}$$

$$h[n] = \begin{cases} 0, & \text{for } n < 0 \\ h[0], & \text{for } n = 0 \\ \varepsilon[n] e^{-\nu(\frac{n}{f_s})}, & \text{for } 1 \leq n \leq N - 1 \end{cases} \tag{5.4}$$

Where *fs* denotes the sampling frequency, T_e refers to the duration for the early reflection part, and *N* indicates the length of the finite impulse response filter. $\varepsilon[n]$ is a random-variable sequence of

independent and identical normal distribution with a mean of zero and a standard deviation of σ_0^2, and the decay factor, v, is determined by the reverberation time, RT_{60}, as

$$v = \frac{3\ \ln(10)}{\mathrm{RT}_{60}} \tag{5.5}$$

The power envelope of $h[\mathrm{n}]$ is equal to:

$$E\{h^2[n]\} = \sigma_0^2 e^{-v(\frac{n}{fs})} \tag{5.6}$$

Where $E\{.\}$ denotes the ensemble averaging. By using (5.2) and (5.4), x[n] can be expressed as the sum of three parts as:

$$x[n] \ = \ \underbrace{h[0]s[n] + \sum_{k=1}^{Te\ fs-1} s[n-k]\ h_e[n]}_{x_e[n]} \ + \underbrace{\sum_{k=Te\ fs}^{N-1} s[n-k]\ h_l[n]}_{x_l[n]} \tag{5.7}$$

The attenuated speech signal from the direct path is $h[0]s[n]$, the reverberated speech from the early reflection part of the impulse response is $x_e[n]$, and the reverberated speech arising from the late reverberation is $x_l[n]$. Alternatively, (5.7) can be expressed as follows:

$$h[0]s[n] + \ x_e[n] = x[n] - \ x_l[n] \tag{5.8}$$

The idea of recovering the original speech signal, $s[n]$ is complicated in terms of algorithm design for human speech recognition. This is attributed to the different perceptual abilities of human listeners compared to machines in utilizing the early reflection part of the signal, $x_e[n]$, to facilitate recognition of reverberated speech [45]. Therefore, the short-time reverberation-suppressed speech signal, $\hat{s}[j,n]$, for frame j can be estimated from the short-time reverberated speech signal, $x[j,n]$, and the short-time late-reverberation signal, $x_l[j,n]$, as:

$$\hat{s}[j,n] = h[0]s[j,n] + x_e[j,n] = x[j,n] - x_l[j,n], \text{ or}$$

$$x[j,n] = x_l[j,n] + \hat{s}[j,n] \tag{5.9}$$

or equivalently in the frequency domain as:

$$\hat{S}(j,p) = X(j,p) - X_l(j,p), \text{ or}$$

$$X(j,p) = X_l(j,p) + \hat{S}(j,p) \tag{5.10}$$

where $\hat{S}(j,p)$, $X(j,p)$and $X_l(j,p)$ denote the Fourier transform at frequency index p of, $s[j,n]$, $x[j,n]$, and $x_l[j,n]$ respectively. Although (5.9) and (5.10) provide a straightforward approach of estimating the clean speech signal, in practice, the instantaneous short-time signal $x_l[j,n]$ [or equivalently $X_l(j,p)$ that is generated by a random process are not accessible. Therefore, $X_l[j,p]$ will have to be estimated from the reverberated speech signal, $x[j,n]$. From (5.9), it is evident that if $x_l[j,n]$ and $\hat{s}[j,n]$ are uncorrelated random signals, then the auto-correlation functions $r_{xx}[j,\tau]$, $r_{x_l x_l}[j,\tau]$ and $r_{\hat{s}\hat{s}}[j,\tau]$ of $x[j,n]$, $x_l[j,n]$ and $\hat{s}[j,n]$, respectively, should satisfy:

$$r_{xx}[j,\tau] = r_{x_l x_l}[j,\tau] + r_{\hat{s}\hat{s}}[j,\tau] \text{ , or}$$

$$r_{\hat{s}\hat{s}}[j,\tau] = r_{xx}[j,\tau] - r_{x_l x_l}[j,\tau] \tag{5.11}$$

Considering the short-time speech signal, $s[j,n]$, can be assumed stationary during a time span between 20 ms and ~40 ms, and that the early and late components, $h_e[n]$ and $h_l[n]$, of the RIR filter are uncorrelated random signals and therefore this seems to be justified approximation [31]. By observing that the reverberated speech signal from late reverberation $x_l[j,n]$ is obtained after the convolution between $s[j,n]$ and the exponentially decayed signal, $h_l[n]$, intuitively, we show

that the auto-correlation function, $r_{x_l x_l}[j,\tau]$, is an exponentially decayed and delayed version of $r_{xx}[j,\tau]$. It can also be shown that if $T_e \ll RT_{60}$,

$$r_{x_l x_l}[j,\tau] = e^{-2\nu T_e}\, r_{xx}[j-(T_e f_s)/R,\tau] \tag{5.12}$$

where *R* is the frame rate in samples. Now, if we consider the autocorrelation functions of (5.11) and (5.12), we can write their power spectral densities as;

$$P_{\hat{s}\hat{s}}[j,p] = P_{xx}[j,p] - P_{x_l x_l}[j,p] \tag{5.13}$$

$$P_{x_l x_l}[j,p] = e^{-2\nu T_e}\, P_{xx}[j-(T_e f_s)/R,p] \tag{5.14}$$

And equation (5.13) is similar to the classic noise reduction method based on spectral subtraction proposed by Boll [6]. In that spectral subtraction strategy relies on obtaining an accurate estimate

$$\left|\hat{S}(j,p)\right| = \max[\,|X(j,p)| - \alpha\left|\hat{X}_l(j,p)\right|,\ \beta\,\left|\hat{X}_l(j,p)\right|\,] \tag{5.15}$$

of the noise spectrum that can be subtracted from the noisy speech spectrum in order to yield a noise-attenuated speech signal. This strategy for noise reduction have been applied to CIs with much success [40]. In the case of reverberant energy, the spectral variance of the estimated clean speech signal, $\hat{s}[j,n]$, can be calculated by subtracting the spectral variance of late reflections $x_l[j,n]$ from that of reverberated speech $x[j,n]$. Although, SS is a very efficient method suitable for real-time implementation, it often suffers from an artifact called "musical noise" due to inaccurate power spectral density estimate. In order to ameliorate this problem, we investigate a method that combines the magnitude spectral over-subtraction and the spectral flooring [5], which showed the most promising results in our pilot study with one CI subject. More specifically, as shown in Figure. 5.2, the magnitude spectrum, $\left|\hat{S}(j,p)\right|$, can be computed by:

where max[.] denotes the absolute maximum value operator, α is the subtraction factor ($\alpha \geq 1$), and β is a small constant for the spectral floor. The subtraction factor, α, is computed as:

$$\alpha = \begin{cases} \alpha_+, & \text{for } SNR_{post} \geq 20 \\ \alpha_0 - SNR_{post}\ s, & \text{for } -5 \leq SNR_{post} \leq 20 \\ \alpha_-, & \text{for } SNR_{post} \leq -5 \end{cases} \quad (5.16)$$

and $$s = \frac{\alpha_- - \alpha_+}{25} \qquad \alpha_- = 20,\ \alpha_+ = 1,\ \alpha_0 = \alpha_+ + 20s \quad (5.17)$$

where SNR_{post} is the *a posteriori* signal-to-noise ratio defined as the ratio of $|X(j,p)|^2$ to $\left|\hat{X}_l(j,p)\right|$. This varying subtraction factor, α, across frames is used to control the trade-off between speech distortion and musical noise. And the constant factor, β, is used to avoid the physically impossible negative magnitude spectral values and account for the case where $|X(j,p)|$ is no greater than $\alpha\left|\hat{X}_l(j,p)\right|$.

Based on the block diagram of Figure. 5.3, $\left|\hat{X}_l(j,p)\right|$ is the estimated magnitude for $|X_l(j,p)|$which can be written as,

$$\sigma_x^2(j,p) = \eta\,\sigma_x^2(j-1,p) + (1-\eta)|X(j,p)^2|\,,$$

$$\left|\hat{X}_l(j,p)\right| = e^{-\nu T_e}\sqrt{\sigma_x^2\left(j - \frac{T_e \cdot f_s}{R}, p\right)} \quad (5.18)$$

where η is a smoothing factor ($0 \leq \eta < 1$) that controls how fast $\left|\hat{X}_l(j,p)\right|$ adapts to the abrupt changes in $|X(j,p)|$.

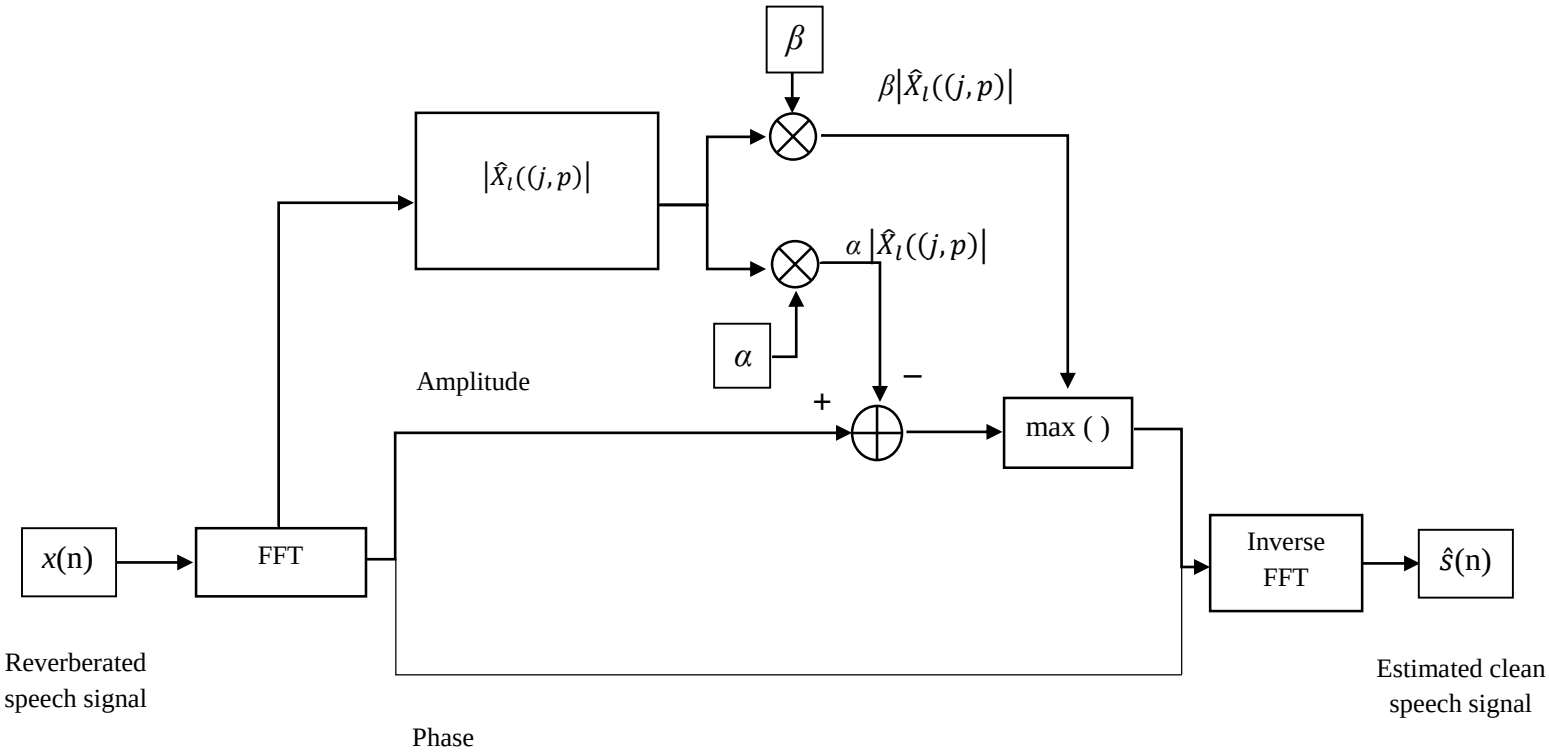


Figure 5.2. Overview of the proposed reverberation suppression strategy based on Spectral Subtraction.

5.5 Implementation of Spectral Subtraction strategy

In this section, we will be discussing the procedure and speech stimuli used for evaluating the proposed SS strategy. Understanding of reverberant speech by CI listeners was assessed with and without the proposed SS strategy.

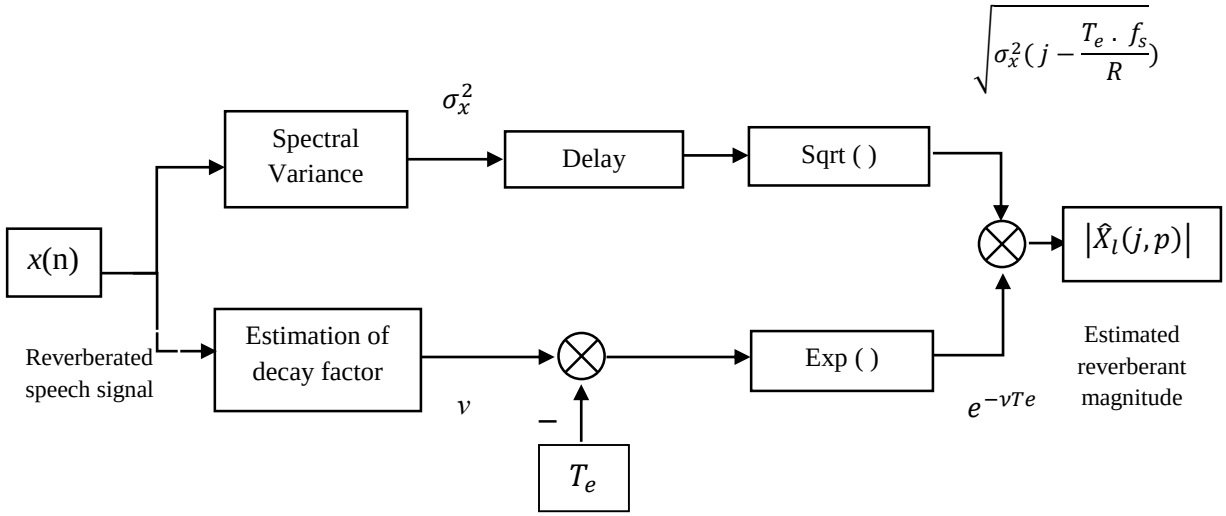


Figure 5.3. Estimation of reverberant magnitude; $|\hat{X}_l(j,p)|$.

5.5.1 Subjective listening tests

Eleven postlingually deafened adult CI listeners were recruited for testing that was conducted at the University of Kansas and the University of Wisconsin–Milwaukee. The listeners were all native speakers of American English and were paid an hourly wage for their participation. All listeners had a minimum of one year of experience with their CI devices, and they used their own processors during testing. Prior to data collection, this study was approved by the Human Subjects Committee of the University of Kansas in Lawrence and the University of Wisconsin–Milwaukee's Human Research Protection Program. A case history interview was conducted with each subject to determine eligibility for this experiment. Only subjects who scored > 70% on the consonant-nucleus-consonant (CNC) test [82] were included in the subject population. The

participants of the study signed a written informed consent prior to the testing. Complete demographic information for the individuals tested is provided in Table 5.1.

The speech stimuli used for testing were sentences from the IEEE database [46]. In total there were 72 lists of 10 sentences, all produced by the same male talker. Speech stimuli were recorded at the sampling frequency of 25000 Hz and were downsampled to 16000 Hz. The root-mean-square value of all sentences was equalized to the same value corresponding to ~65 dBA. Head-related transfer functions (HRTFs) [89] were used to simulate the two acoustically reverberant conditions (RT_{60} = 0.3 s and 1.0 s). The HRTF measurements method is discussed in Appendix B. A CORTEX MKII manikin artificial head was placed inside a rectangular reverberant room with dimensions 5.50 m x 4.50 m x 3.10 m (length x width x height) and a total volume of 76.80 m3 to obtain measurements of the HRTFs. The average reverberation time of the room (averaged in one-third-octave bands with center frequencies between 125 Hz and 4000 Hz) was RT_{60} = 1.0s.

In order to reduce the average reverberation time to RT_{60} = 0.3s, we added extra floor carpeting to the room and highly absorbent rectangular acoustic boards and panels and also increased the number of acoustic panels in the room. This setting corresponds to a well-dampened room with reverberation time often encountered in small office spaces. Identical microphones used in modern behind-the-ear speech processors were used to facilitate the recordings. Finally, the HRTFs obtained for each reverberation condition were convolved with the speech files from the IEEE test materials to generate the corrupted stimuli. This was implemented using standardized linear convolution algorithms in MATLAB (MathWorks, Natick, MA).

5.5.2 Testing procedure

Prior to testing, each subject participated in a practice session to gain familiarity with the task. During the practice session, the listeners adjusted the volume to reach a comfortable level. During the test session, a 10 min break was provided every hour in order to avoid fatigue. Later, the test stimuli were presented to the listeners directly through the auxiliary input jack of the CI sound processor in a double-walled sound-attenuated booth (Acoustic Systems, Inc., Austin, TX).

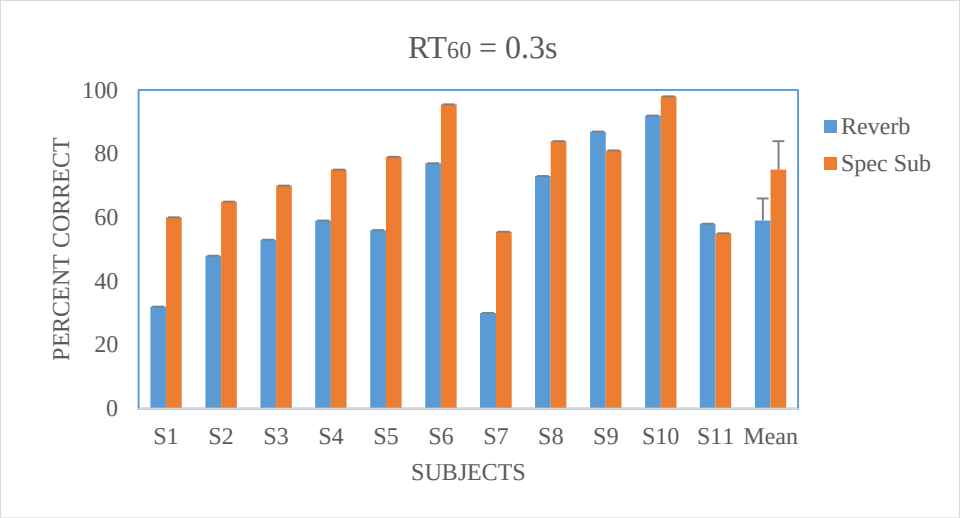
Four experimental conditions corresponding to: (a) two different reverberant conditions (RT_{60} = 0.3 s and 1.0 s) and (b) two processed conditions, one using the proposed spectral-subtraction based reverberation suppression strategy and one using unprocessed reverberant stimuli fed directly into the sound processors. . Two IEEE sentence lists (20 sentences) were used per condition and each sentence was presented once. The order of the test conditions was randomized across subjects to minimize the list effect if any. During testing, participants were allowed to repeat the sentence once and they were instructed to write as many of the words as they could identify on the answer sheets. The responses of each individual were scored off-line based on the number of words correctly identified. All correctly identified words were scored. The percent correct scores for each condition were calculated by dividing the number of words correctly identified by the total number of words in the particular sentence list.

5.5.3 Experimental results

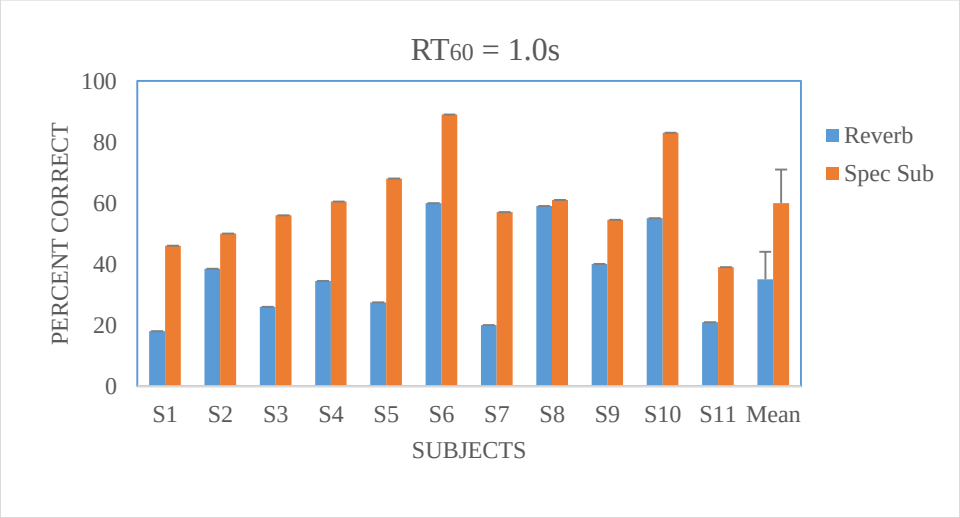
The individual speech intelligibility scores for the two different reverberant conditions (RT_{60} = 0.3 s and 1.0 s) tested are shown in Figures. 5.4 (a) and (b). As it can be seen, significant gains in word recognition scores were obtained in both reverberant stimulus conditions using the proposed SS method. As shown in Fig. 5.4 (a), the mean speech intelligibility scores improved from an average of 59% – 75% for RT_{60} = 0.3 s condition and the intelligibility scores were 25% higher for RT_{60} = 1.0 s when compared to the unprocessed listening condition.

Subject	Gender	Age at testing (Years)	CI experience (Months)	Etiology	CI Device	For SS Strategy	For comparison of IRM and SS
S1	F	30	38	Hereditary	Nucleus 5	X	X
S2	F	49	30	Noise Exposure	Nucleus 5	X	X
S3	M	22	195	Unknown	Nucleus 5	X	X
S4	M	58	65	Hereditary	Nucleus 5	X	X
S5	F	43	24	Unknown	Nucleus 5	X	X
S6	M	67	20	Noise Exposure	Nucleus 6	X	X
S7	M	35	68	Meniere's	AB Harmony	X	X
S8	M	32	120	Meningitis	AB Harmony	X	
S9	M	18	192	Spinal Meningitis	MED-EL	X	
S10	M	66	72	Accident	MED-EL	X	
S11	M	69	120				
Mean		44.5	85.8				
Standard Deviation		18.5	63.3				

Table 5.1. Demographic details of the CI listeners participated in the study.



(a)



(b)

Figure 5.4. Individual percent correct scores for eleven CI listeners tested with IEEE sentences using clean acoustic inputs recorded in reverberation (blue bars) and reverberant acoustic inputs processed with the proposed SS strategy (orange bars), (a) RT_{60} = 0.3 s and (b) RT_{60} = 1.0 s. Standard deviations are indicated by the error bars.

Statistical tests were also performed using two-way analysis of variance (ANOVA) (with repeated measures) with the two different reverberant listening conditions and the two processing conditions (unprocessed and processed using SS) as within-subject factors was conducted on the rationalized arcsine unit-transformed values of the speech perception scores. A Kolmogorov-

Smirnov test was run to confirm the normality of the transformed percent correct scores. Note that a critical value equal to 0.05 was used as the significance level on the statistical analyses performed. The ANOVA indicated a significant effect ($F[1,10] = 46.95$, $p < 0.001$) of the processing strategy, significant effect of reverberation time ($F[1,10] = 60.54$, $p < 0.001$) and significant interaction ($F[1,10] = 10.98$, $p < 0.008$). The observed interaction is due to the fact that the improvement in performance obtained with the proposed SS strategy, relative to the baseline condition, differed for the two reverberation times tested. For RT_{60} = 0.3 s, the mean scores observed in the processed condition were higher than those obtained in the unprocessed (reverberant) condition, albeit with only a marginal statistical significance ($p = 0.047$). On the other hand, the mean scores obtained in the RT_{60} = 1.0 s condition, indicated that speech intelligibility improved significantly ($p = 0.003$) when processing the corrupted stimuli with the SS dereverberation strategy.

5.6 Comparison of Spectral subtraction and Ideal reverberant mask strategies

In this section we will be comparing the improvement in speech intelligibility provided by applying the ideal reverberant mask previously developed by Kokkinakis *et al.* [55] against the improvement observed when processing with the SS strategy introduced in the section 5.3. The IRM threshold in this case was set to -5 dB, as this has been shown to be appropriate for suppressing moderate levels of reverberation. The ideal reverberant (binary) masking strategy has been extensively utilized as a performance upper bound for dereverberation algorithms [34, 35, 55].

5.6.1 Overview of IRM strategy

IRM is a channel selection strategy is based on SRR measurement from the clean and reverberant signals. By comparing the individual channel-specific SRR values against an empirically

determined threshold value the reverberant energies in the gaps can be removed and thus overlap masking effect is reduced. The SRR selection criterion was discussed in Chapter 3. Although this strategy was used in development of algorithms, it relies on the information from clean and is set as a baseline to several algorithms, therefore we use it to compare the spectral subtraction strategy.

5.6.2 Subjective listening tests

All the postlingually deafened cochlear implantees tested with spectral subtraction strategy in section 5.5 were invited back for additional testing but only the first seven CI subjects who were using Nucleus 24 devices were available to participate in this second part of the study (see Table 5.1). All CI listeners tested used their clinical devices programmed with the advanced combination encoder (ACE) speech coding strategy [105]. ACE is the default strategy used in Nucleus devices and operates by selecting only a subset of envelopes (typically between 8 and 12) for stimulation at each cycle. The same speech material from IEEE sentences [46] as in section 5.3 were used here. To avoid potential learning effects, none of the sentence lists previously used were reused.

5.6.3 Testing procedure

In this experiment the subjects participated in a total of four testing conditions corresponding to (a) reverberant stimuli (RT_{60} = 0.3 s and 1.0 s) processed with the IRM strategy described in section 5.2.1, and (b) reverberant stimuli (RT_{60} = 0.3 s and 1.0 s) processed with the proposed spectral-subtraction based reverberation suppression strategy described in section 5.2.2. As before, each participant completed all conditions in a single test session. Two IEEE sentence lists (20 sentences) were used per condition and each sentence was presented once. A total of 80 sentences (20 sentences * 4 test conditions) were used in this experiment. The order of the test conditions were randomized across subjects as before. All the stimuli were presented directly through the auxiliary input jack of the CI sound processors.

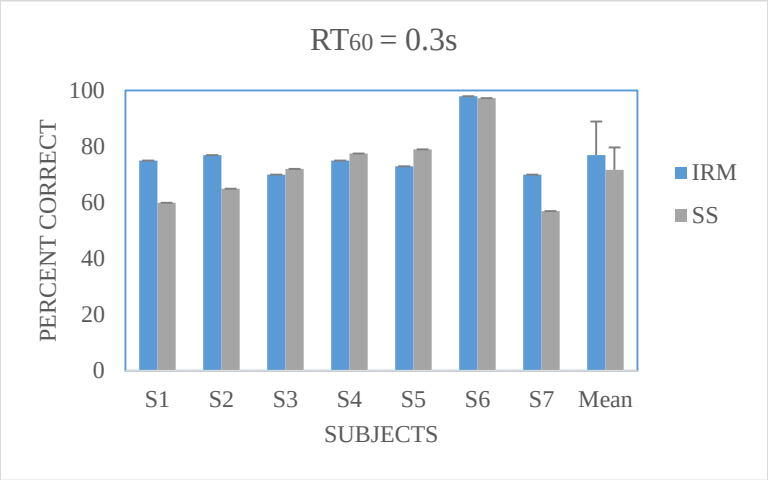
5.6.4 Experimental results and evaluation

Individual subjects' scores obtained using the two aforementioned dereverberation strategies were shown in Figure 5.5. For RT_{60} = 0.3 s, the average speech intelligibility scores for all listeners tested was approximately equal to 77% for the IRM condition and ~72% when the corrupted stimuli were processed using SS. The individual subject scores revealed that almost half of all the individuals tested did better with the SS technique in the 0.3 s condition. For RT_{60} = 1.0 s, the mean scores across all subjects were equal to 70% and 61%for the IRM and SS conditions, respectively. In this condition, the scores indicated that the majority of subjects tested performed better when processing the corrupted stimuli using the IRM strategy. Two-way ANOVA (with repeated measures) using the two different listening conditions and the two processing conditions (IRM- and SS-processed) as within-subject factors was conducted on the rationalized arcsine unit-transformed values of the speech perception scores. The ANOVA indicated a non-significant effect ($F[1,10] = 4.53$, $p = 0.078$) of the processing strategy, a significant effect of reverberation time ($F[1,10] = 45.84$, $p = 0.001$) and a non-significant interaction ($F[1,10] = 2.14$, $p = 0.19$) between the condition and the strategy.

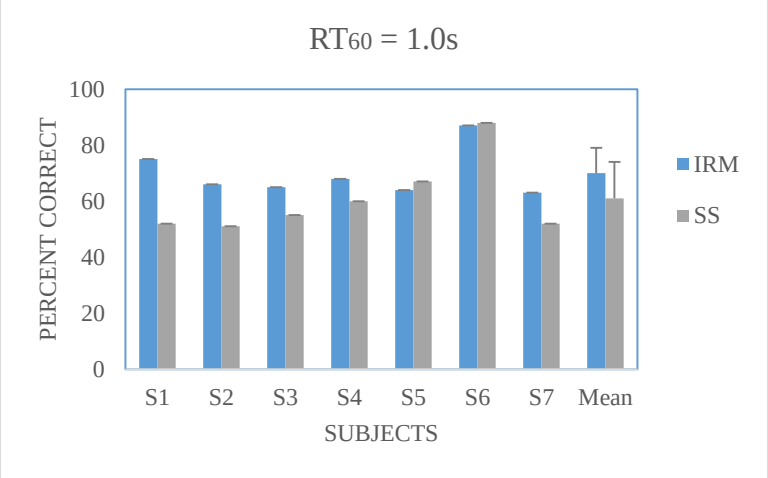
5.7 Summary and Conclusion

The average speech intelligibility scores for all CI listeners tested were obtained for the two reverberant conditions (RT_{60} = 0.3, 1.0 s) which showed improvements with proposed approach. The findings of this study are consistent with the results obtained from an earlier study by Kokkinakis *et al.* [55], which concluded that there is a very strong, and negative, relationship between speech perception and the amount of additive acoustical reverberation. In line with the

study, speech intelligibility by CI users in reverberation degraded exponentially with a linear increase in reverberation time. A number of factors contributed to the low performance observed.



(a)



(b)

Figure 5.5. Individual percent correct scores obtained from seven CI listeners tested with IEEE sentences using reverberant acoustic inputs processed with the IRM strategy (blue bars) and the proposed SS strategy (grey bars), (a) RT_{60} = 0.3 s and (b) RT_{60} = 1.0 s. Standard deviations are indicated by the error bars.

These include the low spectral resolution due to standard CI sound processing, self-masking effects (causing at formant transitions) that are often associated with early reflections, and the detrimental effects of temporal envelope smearing (overlap-masking). Of these three factors, we believe that

the negative effects of temporal envelope smearing, introduced predominantly by overlap-masking, contributed the most. Overlap-masking results in temporal smearing, which often causes one phoneme to be masked by the reverberant energy present in a preceding phoneme. This smearing caused by reverberation (particularly by late reflections) makes the detection of acoustic landmarks extremely difficult, and that disrupts the syllable structure, which is known to be important for determining word boundaries [92].

Also to demonstrate the efficiency of the proposed method, stimulus output patterns (electrodograms) of an IEEE sentence ("The water in this well is a source of good health") uttered by a male speaker and processed with the ACE speech coding strategy were shown in Figure 5.6. In all panels shown, the vertical axes represent the electrode position corresponding to a specific frequency, while the horizontal axes show time progression. In this example, we used speech corrupted by additive reverberation in a room with RT_{60} = 1.0 s. As seen in Figure.5.6 (b), vowel and consonant boundaries were seen to be smeared when compared to the clean stimuli in Figure.5.6 (a). The prolonged formant transitions caused by self-masking are also evident in Figure.5.6 (b) (electrodes 10–15) when focusing on vowel targets (e.g., see vowel /ɔ/ at 0.30–0.45 s and vowel /e/ at 1.10–1.25 s). Such self-masking (or coloration) effects, which arise due to the internal smearing of energy within each phoneme, are believed to be either less or equally damaging to overlap-masking effects [15, 45].

Figure 5.6 (c) depicts the output patterns of stimuli processed using the ideal reverberant mask. As it is evident from the figure that the IRM strategy can suppress additive reverberant energy to a substantial extent. The IRM strategy relies on a local SRR criterion and can correctly discard the reverberant channels corrupted by late reflections, while selecting only those T-F channels containing primarily the signal consisting of the direct component and some early

reflections. As shown in Figure. 5.6 (b), the ACE strategy will select mainly the low- and mid-frequency channels (electrodes 10–18) instead of the high-frequency channels when coding consonant segments. This is due to the principle underlying the operation of ACE, in that speech can be understood well even if only the peaks in the short-term spectrum are transmitted.

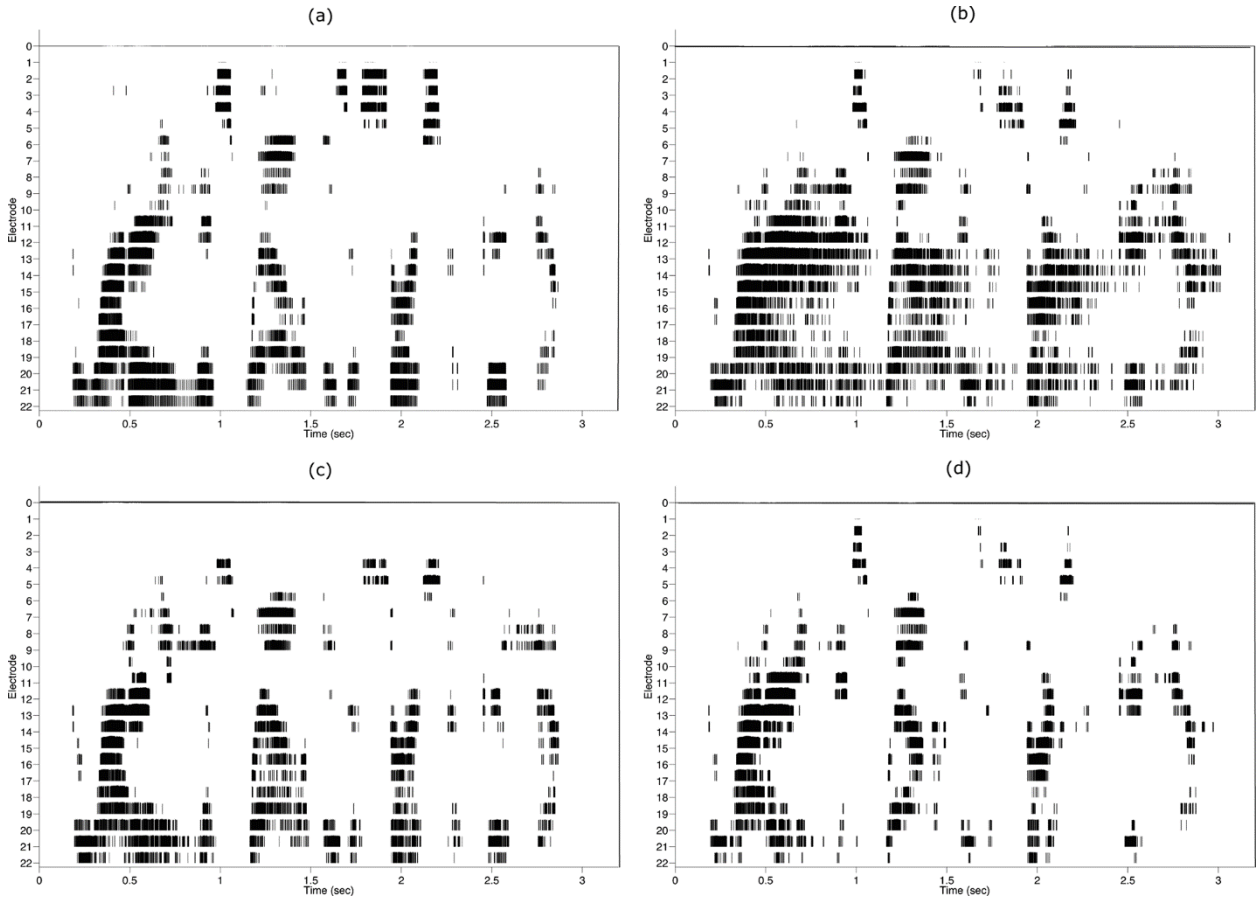


Figure 5.6. Electrodograms of a sentence from the test stimuli. (a) Clean (unmodified) sentence (b) Reverberated stimuli for RT60 = 1.0s (c) Stimuli processed by IRM strategy, and (d) Stimuli processed by SS strategy.

For reverberant stimuli, intense low- and mid-frequency energy present in the stimulus could capture all the available electrode-driving time slots. Thus, none will remain available to respond to the weaker high-frequency energy contained in the unvoiced part of the speech input.

In contrast, the IRM strategy can correctly select the high-frequency channels in the most basal electrodes (electrodes 1–9) corresponding to unvoiced segments (e.g., consonants). Therefore, when pre-processing the corrupted (reverberant) speech stimuli with the IRM strategy, the correct bands with the highest spectra will be selected to drive the electrodes, instead of the mid-frequency channels that contain mainly unwanted additive reverberant energy. As illustrated in Figure. 5.6 (d), low- and mid-frequency energy due to additive acoustic reverberation can also be attenuated considerably (especially in the apical electrodes) after processing with the proposed SS strategy. In addition, the vowel and consonant boundaries, as well as silence gaps previously filled with reverberant energy are recovered resulting in improved useful acoustic cues and an increase in the observed speech intelligibility. In fact, when comparing the electrodogram of the IRM-processed reverberant speech shown in Figure. 5.6 (c) with that of the spectral subtraction-processed reverberant signal in Figure. 5(d), it is evident that both strategies were able to eliminate temporal envelope smearing effects caused by overlap-masking to almost an equal extent. This is also reflected in the non-significant gap observed in the intelligibility scores between the two separate processing conditions. The advantage of the proposed strategy is that it can operate with little to almost no prior information on the signal or the room characteristics, while the IRM strategy requires estimating the SRR criterion from the uncorrupted signal envelopes. This makes the proposed strategy amenable to real-time implementation.

In conclusion, spectral subtraction strategy capable of suppressing reverberant energy components due to late reflections was proposed. Unlike other strategies, this strategy targeted the reverberant energies from late reflection. This strategy was evaluated using CI listeners and two different reverberation time constants and the results were partly published [58]. Speech intelligibility scores observed in the processed conditions suggest that the proposed solution can

be used to successfully ameliorate the negative impact of late reflections on the perceived speech. Given the potential benefits, in terms of intelligibility, such a strategy could be potentially implemented in clinical CI processors for reverberation suppression in complex listening settings.

CHAPTER 6

COMBINING HARMONIC REGENERATION WITH NOISE SUPPRESSION TO IMPROVE SPEECH RECOGNITION IN BACKGROUND NOISE

6.1 Introduction

Despite tremendous advances in cochlear implant technology, background noise continues to be a problem. The concern with the background noise is that it interferes with the ability to hear, understand and differentiate between sounds to be able to concentrate on the target speaker. For cochlear implant listeners, the increased difficulty in understanding speech in background noise is due to the reduced audibility of speech, weakly conveyed F0 cues due to the poor spectral resolution and lack of temporal fine structure. Previous studies on noise reduction showed improvement in speech intelligibility, however they introduce severe distortion in the harmonic structure which consequently degrades the F0 representation [41, 112]. Many research studies have shown that combined electric and acoustic stimulation (EAS) significantly improves speech recognition in noise due to the improved F0 representation of target speech [17, 52, 59, 103, 107]. The combined voice production mechanism produces the variety of vibrations and spectral-temporal compositions that form different speech sounds. Voiced sounds are produced by a repeating sequence of opening and closing of glottal folds with a frequency of between 40 (e.g. for a low frequency gravel male voice) to 600 (e.g. for female children's voice) cycles per second (Hz) depending on the speaker, the phoneme and the linguistic and emotional/expressional context. Temporally and spectrally shaped by the frequency of the openings and closings of the glottal folds. The periodicity of the glottal pulses determines the fundamental frequency (F0) of the laryngeal source and contributes to the perceived pitch of the sound. Several studies on combined

electric and acoustic stimulations have shown that F0 cues are beneficial for improving speech intelligibility in noise. In this chapter, we propose the idea of combining harmonic regeneration of voiced speech segments after noise reduction to further enhance the speech intelligibility. This is supported on the basis of harmonic model [42, 83] and significance of fundamental frequency (F0) and formant information that forms the harmonic structure of speech segments. Therefore, realizing the contributions of harmonics to the benefit of electric and acoustic stimulations, we discuss the importance of regenerating the distorted harmonic structure in the voiced segments after noise is suppressed.

This chapter is organized in the following order. In section 6.2, the motivation of this study to regenerate the harmonics after noise reduction is discussed. To support the need for regenerating the harmonic structure and importance of F0 cues and formant information, Section 6.3 will present the contributions of harmonic information to the benefits of electroacoustic stimulations. In section 6.4, the proposed algorithm of combined noise reduction and harmonic regeneration is presented. Section 6.5 presents two objective measures namely PESQ and STOI that are implemented to test the hypothesis in this study. Analysis of results from objective metrics is presented in Section 6.6 and then finally summary and conclusions are given in Section 6.7

6.2 Motivation

Cochlear implant listeners continue to have difficulty understanding speech and identifying the target speaker in the presence background noise. In many speech recognition and voice communication applications, efficient noise reduction techniques are required to improve the quality of speech. As it can be seen from Figure 6.1, the speech spectrum is degraded and the additive noise corrupts the salient features in the speech. The problem of suppressing the energies from additive noise and enhancing speech quality has been widely studied in the past and is still

an active field of research. Most of the noise reduction methods proposed for cochlear implants use preset optimization criteria and are usually based on pre-processing methods [39, 67, 68, 111]. Although preprocessing approach improves signal-to-noise ratio, it introduce unwanted distortion in the signal and are computationally complex [67] therefore do not allow optimization of algorithms to individual users. In addition, the noise suppressed speech no longer possess the properties mandated by the acoustic process of speech production and does not seem to be synergistically with the existing cochlear implant strategies. In other words, these methods treat all frequency components equally and the harmonic structure in voiced segments and associated sidebands are severely distorted. [83, 112].

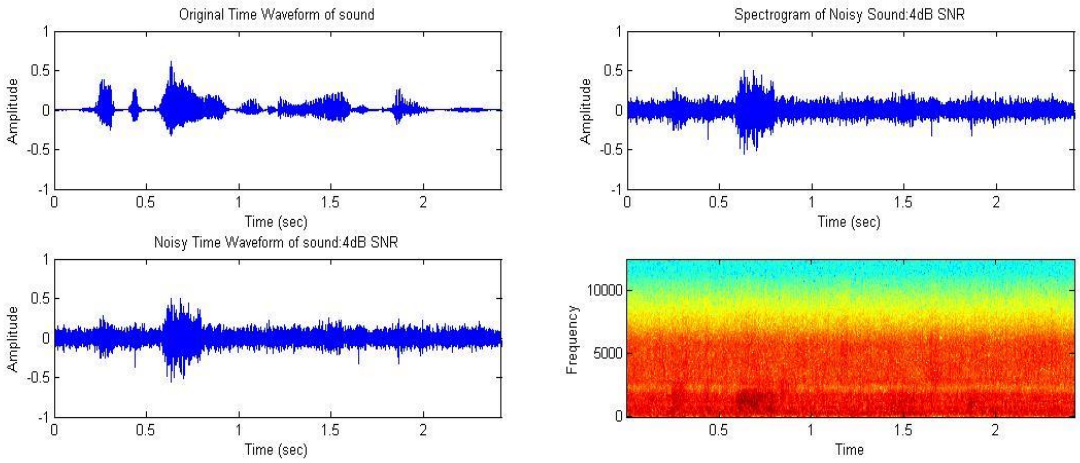


Figure 6.1. Effect of 4 dB SNR noise on clean speech spectrum.

We hypothesize that a deliberately designed post-processing step that regenerates harmonics in the acoustic portion can improve F0 representation, and probably facilitate glimpsing in the voiced region as well. In order to enhance the speech intelligibility after the noise suppression, the distorted harmonics need to be regenerated in the voiced segments as they carry the most of the intelligible information. This study is aimed to integrate the harmonic regeneration technique in noise reduction to further improve the intelligibility of noise-suppressed speech for CI listeners.

Since several studies on electroacoustic stimulations have shown that F0 cues are beneficial to intelligibility, we propose to investigate the synthesis-driven approach to evaluate the contribution of individual and combined harmonics.

6.3 Contributions of harmonics to the benefit of electroacoustic stimulations

6.3.1 Background

Numerous studies have demonstrated significant benefits of combined electroacoustic stimulations for speech perception in noise [17, 52, 59, 103, 107]. Bimodal listeners tested for melody and speech recognition in competing talker backgrounds showed significant improvement in intelligibility [59]. A comprehensive study on speech, voice, and melody recognition by bimodal EAS listeners in quiet and in noise, demonstrated significant improvements [17]. A clinical trial with hybrid cochlear implants showed that EAS listeners gained improved word understanding in noise [28]. More recently, simulated EAS experiments with normal-hearing subjects was introduced to examine the mechanisms underlying the EAS benefits using controlled simulation conditions [9, 62, 88]. From these studies, it was clear that F0 cues present in low-frequency sound helped to group the speech information from high-frequency temporal envelopes, EAS benefits were in part attributed to a better F0 representation and were derived from voicing and glimpsing cues. Synergistic EAS benefits for speech recognition in noise were reported which led to the establishment of mechanisms that extract the low-frequency cues.

Most of the previous studies used lowpass filtering to isolate the F0 regions from the speech stimuli. The extracted segments consisted of only couple of harmonics in the voiced segments and therefore prevents the assessment of other harmonics. Figure 6.2 shows the magnitude spectrum of a typical voiced segment from a male speaker. As illustrated in figure, using lowpass filtering

approach to study EAS benefits is problematic as a cutoff frequency of 250-350 Hz and ignores the remaining harmonics.

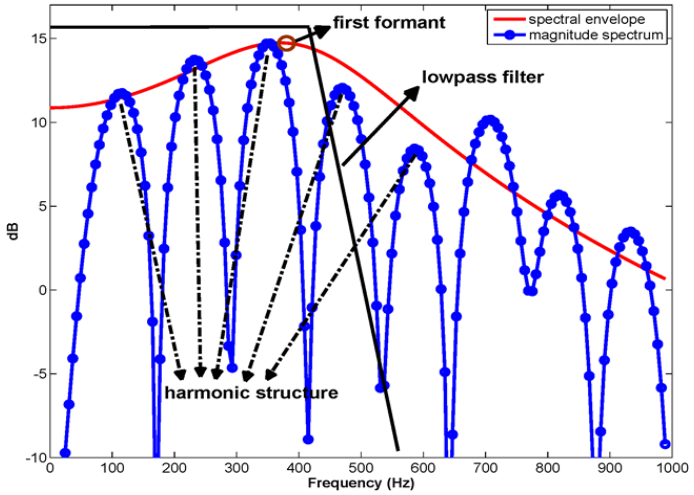


Figure 6.2. Magnitude spectrum of a voiced segment of a male speaker.

The red line is for the spectral envelope, the black solid line represents the lowpass filter fixed for all voiced segments used in previous studies and the red circle indicates the approximate position of the first formant F1. The dashed-dotted line shows the harmonic structure for the voiced speech It can be seen that the generated stimuli contains varying number of harmonics in the voiced speech due to F0 modulation which are not considered in the lowpass filtering approach. This prevents a specific assessment of the contribution of the harmonics to the observed EAS benefits. Therefore we propose a synthesis driven approach, based on harmonic modeling of voiced speech that treat the voiced segments as the sum of a bunch of harmonics. We generate the stimuli by synthesizing sinusoids which means that it is more efficient with implementation of mechanism-driven coding strategies.

6.3.2 Synthesis driven approach for isolating F0 cues

The proposed paradigm is superior to lowpass filtering approach, the reasons are twofold: a better representation of the harmonics improves speech perception by EAS listeners and CI coding

strategies that make harmonics more salient in adverse environment improves speech perception by CI listeners. We synthesized voiced speech using all harmonics above 600 Hz plus different configurations below 600 Hz. This way we were able to precisely control the target cues included in the generated stimuli (see Figure 6.3).

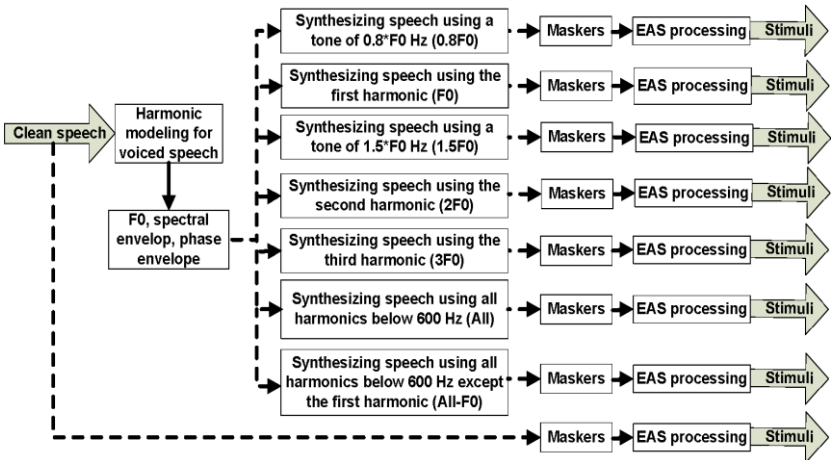


Figure 6.3. Block diagram to generate stimuli with synthesized harmonics for voiced speech segments.

For EAS processing, we synthesized all harmonics above 600 Hz plus several harmonic configurations below 600 Hz. Higher frequency than 600 Hz is vocoded, lower frequency than 600 Hz is lowpass filtered. IEEE sentences from the database [46] were used to test speech in noise tests. Five processing conditions were examined: F0, 2F0, 3F0, All and All-F0. Eight normal hearing listeners were tested and the percent correct scores for two SNR ratios (4dB and 10dB) are shown in the Figure 6.4.

6.3.3 Results of synthesizing harmonics

Results from normal hearing subject indicated that: in addition to F0, other harmonics as well contribute to the EAS benefits. Also, it was observed that under more difficult noisy conditions, higher number of harmonics performed better than conditions with lower number of

harmonics. Hence, it is shown from these results that synthesis driven coding strategies that regenerate harmonics in noise can achieve additional benefits for EAS listeners.

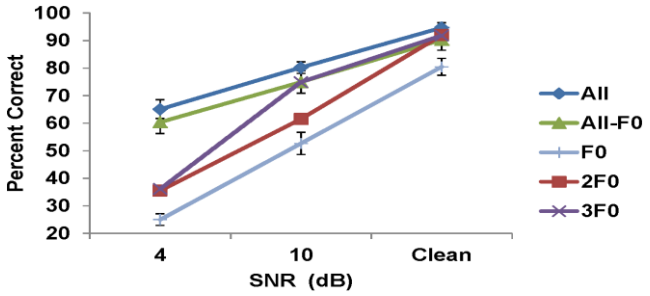


Figure 6.4. Percent correct scores of eight normal hearing listeners tested with synthesized harmonics in EAS condition.

6.4 Combining harmonic regeneration with noise reduction

Under noisy conditions, voiced segments enjoy much higher SNRs compared with unvoiced segments; glimpsing is also easier in voiced segments. Since the majority of speech signals are voiced segments [72], these can be modeled as harmonics and provide prosodic cues. From a speech synthesis point of view, harmonics are easy to generate because the only essential components are F0 and formant structures (F1/F2/F3), and many algorithms exist to estimate both. This makes the harmonic plus noise model (HNM) a popular choice for a speech synthesis system [95]. Harmonics are important in pitch perception and good

In EAS hearing, the low-frequency acoustic portion provides access to several low-numbered harmonics (the harmonic number depends on the F0 value), and this harmonic structure consists of F0 and F1 information. Qin and Oxenham [87, 88] showed that at a cutoff frequency of 300 Hz, it was reasonable to conclude that an improvement in F0 representation aided EAS performance; at a cutoff frequency of 600 Hz, part of the EAS benefits was likely due to improved spectral representation of F1. Hence, preservation of the harmonic structure is critical to F0 representation and likely benefits F1 representation. Although noise reduction improves

glimpsing, but degrades harmonic structure in the process because noise reduction in general does not use any optimization criteria that preserve harmonic structures. To remedy this matter, a post-processing harmonic regeneration step is applied.

The contributions of harmonics to speech intelligibility by cochlear implants showed enhanced F0 representation and improved formant representation. Using the above mentioned synthesis driven approach, it will be beneficial to re-synthesize harmonics in noise reduction. In the context of noise suppression for EAS, we observe identifying voice segments in noisy speech signals will assist in regeneration of harmonics that will improve speech intelligibility.

6.4.1 Algorithm overview

The proposed algorithm is divided into block A: noise reduction and block B: harmonic regeneration as shown in Figure 6.5. The noisy input is first processed through the block A where the fast Fourier Transform extracts the magnitude and phase spectrum. The noise estimation generates the PSD of noise and that is fed to the SNR estimation for a priori and a posteriori SNRs. The suppression gain function is applied to generate noise reduced speech.

For the block B: A simple and efficient way to restore speech harmonics consists of applying a non-linear function; a function that is equivalent to operations of enhancing the signal periodicities, such as *Max*(s(t),0), where s(t) is the noise-reduced speech. However, the harmonic amplitudes of this artificially restored signal are biased compared to clean speech. As a consequence, this signal cannot be used directly as clean speech estimation. Since it contains useful information, a refined a priori SNR can be computed.

Harmonic synthesis algorithms: Adaptive comb filtering, where a comb filter will be used to enhance the harmonic spectral peaks. Training based approach can also be used where distorted

harmonics in the noise-reduced speech will be restored to the associated clean harmonics with models obtained from the training stage.

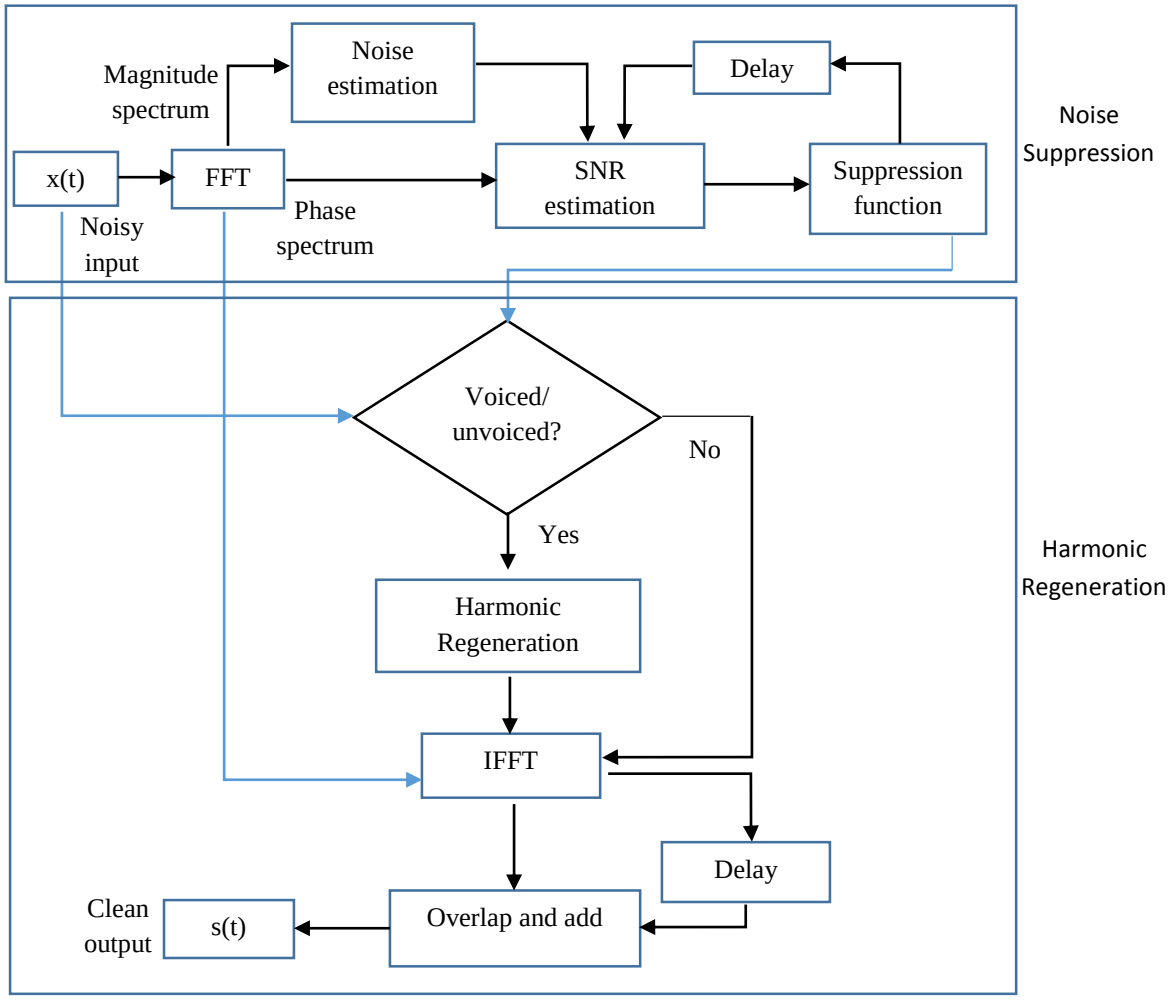


Figure 6.5. Block diagrams for the combined noise suppression (upper block) and harmonics regeneration (lower block).

6.4.2 Algorithm Formulation

For the noise reduction method , the present study is motivated by the two-step noise reduction method developed by Plapous *et al.* [83] which relies on a well-known decision directed (DD) approach developed by Ephraim and Malah [19]. In this approach, we adopted the techniques used

by Plapous *et al.* [83], where the estimation of *a priori* SNR relies on the estimation of the previous frame and hence seems to be biased. Therefore, the two-step noise reduction technique solves this problem refining the estimation of the a priori SNR using a second step to remove the bias of the DD approach, thus removing the reverberation effect. The noise reduction technique relies on the estimation of a short-time spectral gain which is a function of a priori SNR and a posteriori SNR.

In case of additive noise model, if the clean speech $s(t)$ is corrupted by the noise $n(t)$ then the resulting noisy speech is given by $x(t) = s(t) + n(t)$. If $X(j,k)$, $S(j,k)$ and $N(j,k)$ represent the *k*th spectral component of the short-time frame *j* of the noisy speech signal $x(t)$, noise $n(t)$ and clean speech $s(t)$, respectively. In order to enhance the performance of the noise reduction process, the *a priori* SNR is estimated. Using the decision directed approach, the spectral gain $G_{DD}(j,k)$ is computed in the following manner. The goal is to find an estimator $\widehat{S}(j,k)$ which minimizes the expected value of a given distortion measure conditionally to a set of spectral noisy features. The a posteriori and a priori SNR are evaluated for speech enhancement technique. Initially the *a priori* SNR and a posteriori SNR are defined as:

$$SNR_{post}(j,k) = \frac{|X(j,k)|^2}{E\,[|N(j,k)|^2]} \tag{6.1}$$

$$SNR_{prio}(j,k) = \frac{E[|S(j,k)|^2]}{E\,[|N(j,k)|^2]} \tag{6.2}$$

where E[.] denotes the expectation operator. Using noise estimation based on voice detection, the *a priori* SNR and *a posteriori* SNR are computed as;

$$\widehat{SNR}_{post}(j,k) = \frac{|X(j,k)|^2}{\hat{\gamma}_n(j,k)} \tag{6.3}$$

$$\widehat{SNR}_{prio}(j,k) = \beta\,\frac{\left|\hat{S}(j-1,k)\right|^2}{\hat{\gamma}_n(j,k)} + (1-\beta)\,P[\,\widehat{SNR}_{post}(j,k) - 1] \tag{6.4}$$

where P[.] denotes the half-wave rectification and $\hat{S}(j-1,\ k)$ is the estimated speech spectrum at previous frame. This *a priori* SNR estimator corresponds to the so-called decision directed approach whose behavior is controlled by the parameter β (β = 0.98 was used). The spectral gain using the wiener filter is computed as;

$$G_{DD}(j,k) = \frac{\widehat{SNR}_{prio}(j,k)}{1+\widehat{SNR}_{prio}(j,k)} \tag{6.5}$$

In the second step, this gain is used to estimate the *a priori* SNR at frame j + 1.

$$\widehat{SNR}_{prio}(j,k) = \beta\frac{|G_{DD}(j,k)\,X(j,k)|^2}{\hat{\gamma}_n(j,k)} + (1-\beta)\,P[\widehat{SNR}_{post}(j+1,k)-1] \tag{6.6}$$

$$\widehat{SNR}^{SNR2}_{priori}(j,k) = \frac{|G_{DD}(j,k)\,X(j,k)|^2}{\hat{\gamma}_n(j,k)} \tag{6.7}$$

Now, the spectral gain is computed as;

$$G_{SNR2}(j,k) = \frac{\widehat{SNR}^{SNR2}_{priori}(j,k)}{1+\widehat{SNR}^{SNR2}_{priori}(j,k)} \tag{6.8}$$

Finally, the resulting speech spectrum is estimated as follows:

$$\hat{S}(j,k) = G_{SNR2}(j,k)X(j,k) \tag{6.9}$$

This estimate although noise reduced suffers from harmonic distortions and therefore a final stage of harmonic regeneration is applied through the synthesis approach discussed in section 6. 3. A gain function is derived as:

$$G_{har}(j,k) = \frac{SNR_{har}(j,k)}{1+SNR_{har}(j,k)} \tag{6.10}$$

Where

$$SNR_{har}(j,k) = \rho SNR_{har}(j-1,k) + (1-\rho)\max(0,\ \frac{S_{har}(j,k)}{\hat{\gamma}_n(j,k)} - 1) \tag{6.11}$$

And

$$S_{har}(j,k) = |FT(\max(s(j,t),0))|^2 \quad (6.12)$$

Finally, the resulting speech spectrum is estimated as follows:

$$\hat{S}(j,k) = G_{har}(j,k)X(j,k) \quad (6.13)$$

Figure 6.6 shows a reference frame of voiced segment where the clean is represented with a dashed line and the degrade signal is represented with a red colored line. The enhanced speech after harmonically regenerating the harmonics is plotted using the blue line. It appears clearly that some harmonics have been completely suppressed or severely degraded while most are restored using the nonlinear gain function and these results were partly published [97]. In order to measure the performance of the noise reduction and the harmonic regeneration approach used in this study, we chose two intrusive objective measures that have high correlation with subjective tests results.

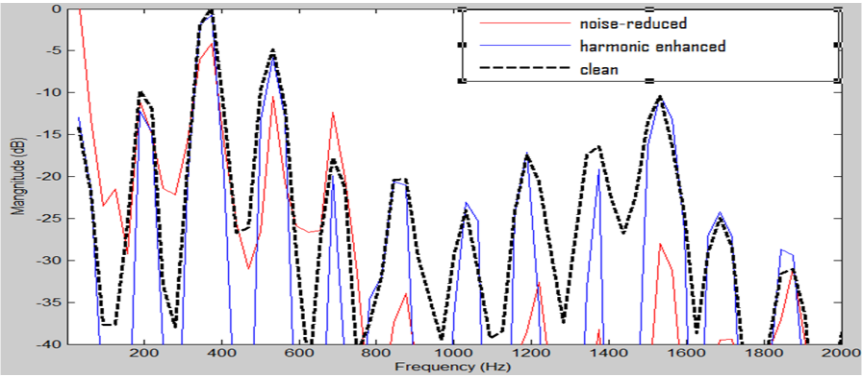


Figure 6.6. Reference frame of voiced segment showing clean, noisy and harmonically enhanced signals.

6.5 Objective measures for speech quality

While subjective listening tests are considered to be most accurate and preferable for evaluation of speech quality, there is excessive amount of time and cost associated with these tests. To resolve

this issue, researchers have adopted an alternative path to develop objective measures that help in predicting the speech quality. As a result of tremendous research in this field, many objective measures have been proposed over the years [1, 48, 69, 93, 96]. Objective metrics allow for setting different system parameters to be tested for efficient, reliable, allows for concurrent processing and potentially play a major role in the development of speech enhancement techniques (noise suppression, dereverberation) for CI devices. Experiments with noisy stimuli showed that majority of the tested measures could accurately predict CI user intelligibility.

Two intrusive objective metrics were evaluated in this study: Perceptual Evaluation of Speech Quality (PESQ) and Short-time objective intelligibility measure (STOI).

6.5.1 Perceptual Evaluation of Speech Quality (PESQ)

PESQ is the International Telecommunications Union (ITU-T) P.862 Recommendation for speech quality assessment of narrow-band and wide-band speech [48]. PESQ can be used to compare the original signal *X(t)* and the degraded signal *Y(t)* to compute the speech quality. The output of PESQ is a prediction of the perceived quality that would be given to *Y(t)* by subjects in a subjective listening test. PESQ performs series of processing including level alignment, input filtering, time alignment and auditory transform that transforms the reference and degraded signals to an internal representation that is analogous to the psychophysical representation of audio signals in the human auditory system, taking account of perceptual frequency (Bark) and loudness (Sone). Following are stages of processing involved in this process: time and level alignment to a calibrated listening level, mapping of time and frequency, frequency warping, and compressive loudness scaling. A detailed overview of this algorithm is explained in Appendix C.1. The algorithm is based on the sensory model that calculates the two disturbance parameters using non-linear averages over specific areas of the error surface: absolute disturbance (Dind) which is a symmetric measure of

absolute audible error and additive disturbance (Aind) disturbance which is an asymmetric measure of audible errors that are much louder than the reference. These disturbances are estimated through a comparison of the clean and processed signals, both mapped to a psychoacoustically-relevant domain. The final PESQ score is a linear combination of the average disturbance value and the average asymmetrical disturbance value using optimized coefficients. Therefore the mean opinion score between 1.0 and 4.5 for the listening quality, which are normally found in an ACR experiment.

$$PESQ = a_0 + a_1.D_{ind} + a_2.A_{ind} \tag{6.14}$$

It was shown that there is a high correlation between the objective scores for PESQ [48], for 22 known ITU benchmark experiments, the average correlation was 0.935. And for an agreed set of eight experiments used in the final validation – experiments that were unknown during the development of PESQ – the average correlation was also 0.935. Therefore, PESQ is an effective measure for speech quality.

6.5.2 Short-time objective intelligibility measure (STOI)

This short-time analysis method [96] is a function of the clean and processed speech, denoted by *x* and *y*, respectively. First, a TF-representation is obtained by segmenting both signals into 50% overlapping, Hanning-windowed frames with a length of 256 samples, where each frame is zero-padded up to 512 samples and Fourier transformed. Then, a one-third octave band analysis is performed by grouping DFT-bins. A TF-unit is then computed for clean and degraded speech. The intermediate intelligibility measure for one TF unit depends on a region of *N* consecutive TF-units from both $X_j(n)$ and $Y_j(n)$. A local normalization procedure is applied, by scaling all the TF-units from degraded unit $Y_j(n)$ with a factor α such that its energy equals the clean speech energy, within that TF-region. Then, $\alpha Y_j(n)$ is clipped in order to lower bound the signal-to-distortion ratio

(SDR). A more detailed explanation of this model is provided in Appendix C.2. The intermediate intelligibility measure is defined as an estimate of the linear correlation coefficient between the clean and modified processed TF-units, which is then averaged over all the frames and bands to obtain the OIM.

6.5.3 Speech material

The target speech materials consisted of sentences from the IEEE database [46]. The IEEE corpus has 72 lists of ten phonetically balanced sentences of high difficulty. The sentences were produced by a male speaker and recorded in a double walled sound-attenuation booth at a sampling rate of 25,000 Hz. Speech shaped noise is used as noise masker as it has been shown to be a very effective masker [62, 103], and its stationarity minimizes the confounding effect of the accuracy of noise estimation algorithms. The experiments were designed to evaluate the benefits of noise reduction and harmonics regeneration in noise; two processing conditions were used for this purpose. Two SNR levels (4dB and 10dB) have been used to test the hypothesis.

6.6 Analysis of results

The performance of each objective metric was evaluated on a per-condition and a per-sample basis. In per-condition case, performance measures were evaluated by averaging the objective ratings for each conditions. In this study, two SNR conditions were present in the noise category (4 dB SNR and 10 dB SNR). In the per-sample case, 72 lists with each 10 sentences were used per SNR level available per degradation scenario (720 sentences × 2 conditions). Two objective measures were evaluated using these stimuli. The mean opinion scores for the PESQ metrics and the percent correct scores for the two SNR levels are shown in Figure. 6.7, where ‘Noproc’ represent the comparison between the clean and the noisy speech, ‘NR’ represent the comparison between the

clean and the noise reduced speech after the application of noise reduction technique and 'HR' represent the comparison between the clean and the harmonically regenerated speech. Performance was measured in terms of the mean opinion scores for PESQ method and it was shown that the scores were higher for 10 dB compared to 4 dB SNR level. As it can be seen that the noise reduction and harmonic regeneration both showed improvement at 10 dB SNR level. The 'NR' scores were computed using PESQ by considering clean as reference and noise reduced as degraded. There was about 10% improvement at 10 dB SNR and about 3% improvement at 4 dB SNR. Similarly, 'HR' scores were computed by considering clean as reference and harmonically regenerated signal as degraded. There was about 14% improvement at 10 dB SNR and about 6% improvement at 4 dB SNR.

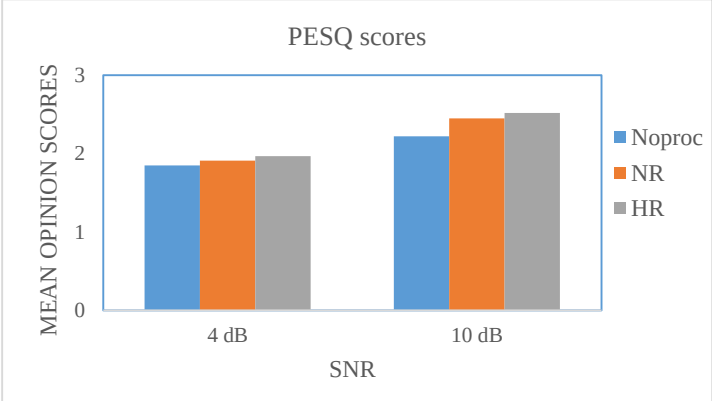


Figure 6.7. Mean opinion scores for PESQ objective metric for two SNR conditions (4dB and 10dB).

On average, the harmonic regeneration approach shows improvement after noise suppression. Multiple paired comparisons with Bonferroni correction were computed between Noproc, NR and HR scores at the both SNR levels. The Bonferroni-corrected statistical significance level was set at $p<0.0125$ ($\alpha = 0.05$). The comparisons indicated statistically significant differences between the Noproc and NR scores ($p<0.006$) and Noproc and HR scores ($p<0.007$). Therefore suggesting that harmonic regeneration can benefit noise reduction in steady-

state noise. The scores obtained with the NR and HR stimuli at lower SNR level were lower, suggesting that EAS conditions can introduce the low frequency information to enhance the harmonics structure. Although this method is the most accurate for stationary noise, there is a disadvantage of complex computations involved in the measurements.

Another objective measure is explored in order to provide the support for the results obtained from PESQ measure. Short-time objective intelligibility measure (STOI) was used to test the intelligibility and quality of the speech processed with the proposed approach. The objective measure produces a correlation coefficient which shows the correlation between the two signals applied to the algorithm [96].

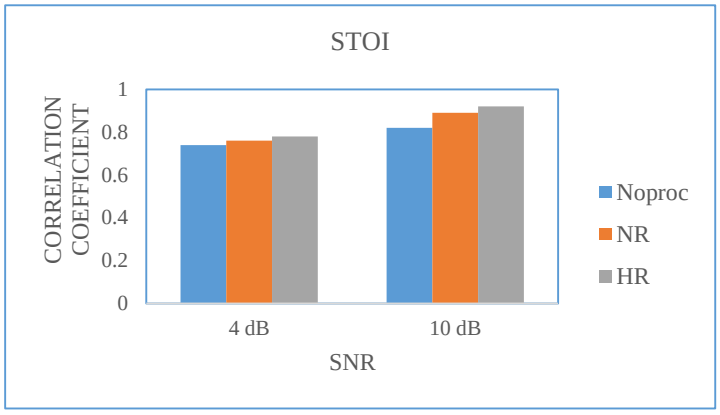


Figure 6.8. Correlation coefficient for STOI objective metric for two SNR conditions (4dB and 10dB).

The correlation coefficients between clean with Noproc, NR and HR scores were evaluated for the two SNR levels are shown in Figure. 6.8. Similar results were obtained from this measure, the lower SNR level (4 dB) performed worse than 10 dB. For 10 dB condition, the correlation between clean and the NR showed 9% improvement and the correlation between the clean and the HR condition showed 12% improvement. The 4 dB condition did not perform well showing little to no benefit. As mentioned in [96], this method is unable to improve intelligibility of noisy in

case of speech-enhancement as it is it is based on an intermediate intelligibility measure for short-time (approximately 400 ms) TF-regions, and uses a simple DFT-based TF-decomposition.

6.7 Summary and Conclusions

Two simple objective intelligibility measures were used to evaluate the performance of proposed approach for noise reduction. The results indicate that the two step noise reduction method provided significant speech intelligibility improvement in steady-state noise. A mean advantage of 10 percentage points was obtained with harmonic regeneration method over the unprocessed condition across the two SNR levels. This outcome is most likely attributed to the two steps involved in the noise reduction technique. On the other hand, noise reduction algorithms degrade harmonic structures, and, as a result, make F0 representation poorer. This is especially true at lower SNR levels, as the impact of inaccurate noise estimation is more significant. As explained in Section 6.3, to overcome this disadvantage, harmonics regeneration can be integrated into noise reduction to improve F0 representation. From both objective metric results, harmonics regeneration provided an additional advantage over the noise reduction at higher SNR levels, although very little benefits were observed at the lower SNR levels as more accurate noise estimation is required for lower frequencies. In frequency domain, from a signal approximation perspective, harmonics regeneration combined with noise reduction align the processed signals better with the speech production model; while in temporal domain, temporal fine structures that carry periodicity information are partly restored. It is expected that the combining the electric and acoustic stimulations (EAS) can benefit the performance of the proposed approach. The accompanying benefits could be improved F0 (fundamental frequency) and F1 (formant) representations.

In this chapter, noise reduction technique based on DD approach was used that relied on the estimation of the *a priori* SNR in two steps. With two steps involved in the processing: *a priori* SNR estimation and removal of frame delay this technique showed the ability avoid any musical noise and be able to track the non-stationarity of the speech signal. Consequently, the speech onsets and offsets are preserved and the reverberation effect characteristic of the DD approach is removed. To compute the spectral gain that preserves the harmonics, the *a priori* SNR is refined through the resulting signal which helps avoids distortions. The role of the nonlinearity and the principle of harmonic regeneration have been detailed and analyzed. Results are given in terms of PESQ and STOI objective metrics on large corpus of 720 sentence lists showed the efficiency of the harmonic regeneration technique. In order to provide complete results of the hypothesis, results from a formal subjective test with cochlear implant patients are to be conducted that confirm the significant performance improvement by the proposed technique. Also combined electroacoustic stimulations could be tested to show improvement in lower SNR levels.

CHAPTER 7

CONCLUSIONS

This chapter concludes this dissertation with a summary of our contributions and a brief discussion on the future direction of this research.

This dissertation assessed various methods to enhance the speech perception of cochlear implant listeners under different adverse environments and proposed new strategies for speech processed by telephone networks, speech in acoustic reverberant conditions and speech in background noise. A significant effect of extending the bandwidth of telephone processed speech was observed in this study. Four filtering configurations: wideband speech (WB), bandpass-filtered speech (BP), highpass-filtered speech (HP) and lowpass-filtered speech (LP) were used to assess the relative importance of adding low and high frequency information to band-limited telephone speech. Experiments with bimodal listeners showed significant improvement in intelligibility when higher frequency information is restored in frequency-limited telephone speech. Interestingly, paired comparison test results showed that LP scores were statistical similar to BP scores and WB scores were statistically similar to HP scores. HP scores were significantly better than LP scores for HA, CI and HA+CI conditions, thus supporting the extension of bandwidth towards the higher frequencies.

In addition, benefits of bimodal listening were also evaluated and compared with CI alone listening in this study. It was worthwhile to examine whether adding hearing aids provide any benefits to recognition of band-limited telephone speech by CI listeners. The results showed that low frequency cues are not useful for CI listening due to the lack of F0 representation in electrical hearing but the addition of contralateral HA showed significant benefit for CI listening.

Most of the devereberation strategies developed for cochlear implants were focused on either reverting the process of reverberation or recovering the direct sound from the reverberated signal. Due to the fact that reverberated sound in an enclosed space consists of direct sound, early and late reflections, it is critical to know that they produce different perceptual effects on the human auditory system and that the reverberation suppression strategies must be able to differentiate between these types of reverberated speech signals. Earlier studies provided strong evidence that the late reverberation are detrimental to overall speech intelligibility compared to the impact of early reflections. In this dissertation, we proposed a reverberation suppression strategy based on spectral subtraction that targets the late reflections by suppressing its reverberant energy components. Spectral subtraction strategy was proven to be efficient for noise reduction and estimating the magnitude of underlying clean speech by subtracting the noise magnitude spectrum from the noisy speech. We investigated whether this strategy could be beneficial to CI listeners by modeling the overlap masking effects caused by late reflections as uncorrelated random noise. Head-related transfer functions (HRTFs) was used to simulate the two reverberant conditions of RT_{60} = 0.3s and 1.0 s. Experiments with cochlear implant patients were conducted for these two reverberant conditions and the performance was measured in terms of percent correct words identified by the CI listener. The results showed significant gains in word recognition scores for both the reverberant stimulus conditions using the spectral subtraction method. The average speech intelligibility scores for CI listeners improved by about 15 percentage points for RT_{60}= 0.3s and about 25 percentage points for RT_{60}= 1.0s reverberant conditions, respectively. These findings strongly prove that additive reverberant energies from late reverberation negatively impacts the speech perception by CI listeners. It was observed that as the reverberation time increased the speech intelligibility decreased exponentially which was consistent with the previous studies and

was attributed mainly to the temporal envelope smearing caused by the overlap masking. With the proposed strategy, low and mid frequency energy due to additive acoustic reverberation are attenuated. In addition, this strategy helps improve intelligibility and useful acoustic cues by recovering the vowel and consonant boundaries, as well as silence gaps that were previously filled with reverberant energy. Thus eliminating the temporal envelope smearing effects caused by overlap masking effect.

Another test was designed to compare the improvement in speech intelligibility provided by applying spectral subtraction strategy against improvements with ideal reverberant masking. Experiments with cochlear implant patients were conducted and the results for spectral subtraction strategy were comparable to the ideal reverberant masking. While the ideal reverberant masking strategy requires estimating the signal-to-reverberant ratio criterion from the uncorrupted signal envelopes, spectral subtraction strategy proposed in this dissertation can operate with little to almost no prior information on the signal or the room characteristics. Due to this benefit, this strategy could be implemented in real-time CI processors for reverberation suppression in adverse listening conditions.

It is well known that listeners with hearing loss have great difficulty perceiving speech in the presence of background noise. Several noise reduction methods were proposed over the years which were based on preprocessing approach to enhance the speech perception in noise. We proposed a method that relies on harmonic regeneration after noise reduction to further enhance the voiced segments of speech. In this study, we evaluated the contribution of harmonics by synthesizing the voiced segments of the speech which proved the importance of harmonic structure to speech intelligibility. The noise reduction algorithm based on two step noise reduction which relies on decision directed approach for a priori SNR estimation was used followed by harmonic

regeneration. The algorithm was evaluated by means of objective intelligibility measures (OIM). Two intrusive objective measure were used namely, Perceptual evaluation of speech quality (PESQ) and Short-time objective intelligibility measure (STOI). For speech shaped noise, results indicate that the regenerated harmonics improved speech intelligibility. It was observed that higher SNR condition (SNR = 10 dB) performed better for noise reduction and harmonic regeneration. Both the objective measures showed the efficiency of the combined harmonic regeneration technique and noise reduction.

7.1 Major contribution of this dissertation

In this dissertation, we proposed novel coding strategies and techniques that improve the speech perception of CI listeners in adverse environments.

- An efficient strategy for assessing the effects of adding low and high frequency information to telephone speech was proposed that provided support for the design of algorithms that would extend higher frequency information at least in quiet environments. It was also shown that the low-frequency cues that are not useful for cochlear implant listeners are actually beneficial when provided to the bimodal listeners via hearing aids.
- A novel reverberation suppression strategy based on spectral subtraction was proposed to suppress the reverberant energy caused by late reflections. The proposed strategy can operate with little to no prior information on the signal and the room acoustics and this makes it amenable to real-time implementation in clinical CI processors.
- A state of the art mechanism underlying the contribution of harmonics to the benefits of electroacoustic stimulations was proposed to enhance the speech intelligibility of cochlear implant users. Synthesis-based approach in the proposed method where beneficial acoustic cues are proven to be more salient for CI listeners.

7.2 Future Research

Future research can be pursued in the following directions:

- Telephone processed speech: The proposed experiments efficiently assessed the contributions of low- and high- frequencies to the band-limited telephone speech and showed that added high-frequency information benefits the telephone speech perception in quiet environments, for both CI and bimodal listeners. Based on these results an efficient bandwidth extension technique that would extend higher frequency information can be designed in future. Also, the current work did not address noisy conditions, and this is a very important issue when noise is introduced in the communication lines. With the use of hearing aids, it is expected that extending frequencies below 300 Hz will facilitate improved representation of F0 cues and glimpsing [62, 114] to benefit bimodal users. Strategies can be developed in future that will evaluate noisy conditions for the telephone processed speech.
- Reverberated speech: Since the spectral subtraction method showed impressive results, more realistic environments with both reverberation and noise need to be examined in future.
- Speech in background noise: Future work is needed to perform subjective listening tests to confirm the results from objective measures on speech quality. Other types of noise maskers could also be tested. Also efforts can be made to obtain better noise estimates in lower SNR conditions that are needed to improve the performance of noise reduction and better F0 estimation for harmonic regeneration.

REFERENCES

[1] ANSI S3.5, "Methods for the Calculation of the Speech Intelligibility Index," *American National Standards Institute*; New York, 1997, Reaffirmed: 2007.

[2] Arndt, P., Staller, S., Arcaroli, J., Hines, A. and Ebinger, K., "Within-subject comparison of advanced coding strategies in the Nucleus 24 Cochlear Implant," *Cochlear Corporation*, 1999.

[3] Assmann, P. F. and Summerfield, Q., "The perception of speech under adverse acoustic conditions," in *Speech Processing in the Auditory System*, edited by Greenberg, S., Ainsworth, W. A., Popper, A. N. and Fay, R. R. (Springer, New York), pp. 231-308, 2004.

[4] Barron, M. and Marshall, A. H., "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure," *Journal of Sound and Vibration*, vol. 77, pp. 211–232, 1981.

[5] Berouti, M., Schwartz, R. and Makhoul, J., "Enhancement of speech corrupted by acoustic noise," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 208-211, 1979.

[6] Boll. S. F., "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transaction on Acoustics, Speech, Signal Processing*, vol. 27, pp. 113-120, 1979.

[7] Bolt, R. H. and MacDonald, A. D., "Theory of speech masking by reverberation," *Journal of the Acoustical Society of America*, vol. 21, no. 6, pp. 577-580, 1949.

[8] Bradley, J. S., Sato, H., and Picard, M., "On the importance of early reflections for speech in rooms," *Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233-3244, 2003.

[9] Brown, C. J., Geers, A., Herrmann, B., Kirk, K. I, Tomblin, J. B., Waltzman, S., Levinson, R., Linn, G. and Brannen, S., "Technical Report: Cochlear Implants," *American Speech-Language-Hearing Association*, TR2004-00041, 2003.

[10] Chang, J. E., Bai, J. Y. and Zeng, F., "Unintelligible low-frequency sound enhances simulated cochlear-implant speech recognition in noise," *IEEE Transactions on Biomedical Engineering*, vol. 53, pp. 2598–2601, 2006.

[11] Cheng, C. I. and Wakefield, G. H., "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time," *Journal of Audio Engineering Society,* vol. 49, no. 4, pp. 231-249, 1999.

[12] Cohen, N. L., Waltzman, S. B. and Shapiro, W. H., "Telephone speech comprehension with use of the nucleus cochlear implant," *The Annals of Otology, Rhinology, and Laryngology Supplement*, vol. 142, pp. 8–11, 1989.

[13] Cray, J. W., Allen, R. L., Stuart, A., Hudson, S., Layman, E. and Givens, G. D., "An investigation of telephone use among cochlear implant recipients," *American Journal of Audiology*, vol. 13, pp. 200–212, 2004.

[14] Desmond, J. M., Collins, L. M. and Throckmorton, C. S., "Using Channel specific statistical models to detect reverberation in cochlear implant stimuli," *Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 1112 – 1120, 2013.

[15] Desmond, J. M., Collins, L. M. and Throckmorton, C. S., "The effects of reverberant self- and overlap-masking on speech recognition in cochlear implant listeners," *Journal of the Acoustical Society of America*, vol. 135, pp. EL304-EL310, 2014.

[16] Dorman, M. F., Hannley, M., Dankowski, K., Smith, L. and McCandless, G., "Word recognition by 50 patients fitted with the Symbion multi-channel cochlear implant," *Ear Hear.*, vol. 10, pp. 44-49, 1989.

[17] Dorman, M. F., Gifford, R. H., Spahr, A. J. and McKarns, S. A., "The benefits of combining acoustic and electric stimulation for the recognition of speech, voice and melodies," *Audiology & Neurotology*, vol. 13, pp. 105–112, 2008.

[18] Drullman., "Temporal envelope and fine structure cues for speech intelligibility." *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 585-592, January 1995.

[19] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transaction on Acoustic, Speech, Signal Process*, vol. 32, pp. 1109-1121, 1984.

[20] Fetterman, B. L. and Domico, E. H., "Speech recognition in background noise of cochlear implant patients," *Otolaryngology-Head and Neck Surgery*, vol. 126, pp. 257-263, 2002.

[21] Fröhne-Büchner, C., Büchner, A., Gärtner, L., Battmer, R. D. and Lenarz, T., "Experience of uni- and bilateral cochlear implant users with a microphone positioned in the pinna," *International Congress Series*, vol. 1273, pp. 93-96, 2004.

[22] French, N. R. and Steinberg, J. C., "Factors governing the intelligibility of speech sounds," *Journal of the Acoustical Society of America*, vol. 19, pp. 90–119, 1947.

[23] Fretz, R. J. and Fravel, R. P., "Design and function: a physical and electrical description of the 3M House cochlear implant system," *Ear Hear 6*, Suppl: 14S-19S, 1985

[24] Friesen, L., Shannon, R., Baskent, D. and Wang, X., "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," *Journal of the Acoustical Society of America,* vol. 110, no. 2, pp. 1150-1163, 2001.

[25] Fu, Q. J., Shannon R. V. and Wang, X., "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," *Journal of the Acoustical Society of America*, vol. 104, no.6, pp. 3586-3596, 1998.

[26] Fu, Q. J. and Galvin, J. J., "Recognition of simulated telephone speech by cochlear implant users," *American Journal of Audiology*, vol. 15, pp. 127–132, 2006.

[27] Gantz, B., Tyler, R. S., Abbas, P. J., Tye-Murray, N., Knutson, J. F., McCabe, B. F., Lansing, C., Brown, C. J., Woodworth, G., Hinrichs, J. and Kuk, F. K., "Evaluation of five different cochlear implant designs: audiologic assessment and predictors of performance," *Laryngoscope,* vol. 98, pp. 1100–1106, 1988.

[28] Gantz, B. J., Turner, C. W. and Gfeller, K. E., "Acoustic plus electric speech processing: Preliminary results of a multicenter clinical trial of the Iowa/nucleus hybrid implant," *Audiology & Neurotology*, vol. 11, pp. 63–68, 2006.

[29] Gifford, R. H., Dorman, M. F., McKarns, S. A. and Spahr, A. J., "Combining electric and contralateral acoustic hearing: word and sentence recognition with bimodal hearing," *Journal of Speech, Language and Hearing Research*, vol. 50, no. 4, pp. 835–843, 2007.

[30] Griesinger, D., "The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces," *Acta Acustica*, vol. 83, pp. 721-731, 1997.

[31] Habets, E. A. P., "Single- and multi-microphone speech dereverberation using spectral enhancement," PhD dissertation, Technische University Eindhoven, Eindhoven, The Netherlands, 2007.

[32] Hamacher, V., Doering, W., Mauer, G., Fleischmann, H. and Hennecke, J., "Evaluation of noise reduction systems for cochlear implant users in different acoustic environments," *American Journal of Otolaryngology*, vol. 18, pp. S46-S49, 1997.

[33] Hazrati, O. and Loizou, P. C., "The combined effects of reverberation and noise on speech intelligibility by cochlear implant users," *International Journal of Audiology*, vol. 51, pp. 437-443, 2012.

[34] Hazrati, O., Lee, J. and Loizou, P. C., "Blind binary masking for reverberation suppression in cochlear implants," Journal *of the Acoustical Society of America*. vol. 133, pp. 1607-1614, 2013

[35] Hazrati, O. and Loizou, P. C., "Reverberation suppression in cochlear implants using a blind channel-selection strategy," *Journal of the Acoustical Society of America*, vol. 133, pp. 4188-4196, 2013.

[36] Hillenbrand, J., Getty, L., Clark, M. and Wheeler, K., "Acoustic characteristics of American English vowels," *Journal of the Acoustical Society of America,* vol. 97, pp. 3099-3111, 1995.

[37] Horng, M. J., Chen, H. C., Hsu, C. J. and Fu, Q. J., "Telephone speech perception by Mandarin-speaking cochlear implantees," *Ear and Hearing*, vol. 28, pp. 66S–69S, 2007.

[38] Hu, Y. and P. Loizou., "A generalized subspace approach for enhancing speech corrupted with colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334-341, 2003.

[39] Hu, Y. and Loizou, P., "Subjective comparison and evaluations of speech enhancement methods," *Speech Communication*, vol. 49, pp. 588-601, 2007.

[40] Hu, Y. and Loizou, P., "A comparative intelligibility study of single-microphone noise reduction algorithms," *Journal of the Acoustical Society of America*, vol. 122, pp. 1777-1786, 2007.

[41] Hu, Y. and Loizou, P., "A new sound coding strategy for suppressing noise in cochlear implants," *Journal of the Acoustical Society of America*, vol. 124, pp. 498-509, 2008.

[42] Hu, Y., "A simulation study of harmonics regeneration in noise reduction for electric and acoustic stimulation," *Journal of the Acoustical Society of America*, vol. 127, pp. 3145-3153, 2010.

[43] Hu, Y. and Loizou, P., "Effects of introducing low-frequency harmonics in the perception of vocoded telephone speech," *Journal of the Acoustical Society of America*, vol. 128, pp. 1280-1289, 2010.

[44] Hu, Y., Tahmina, Q., Runge, C. and Friedland, D. R., " The perception of telephone-processed speech by combined electric and acoustic stimulation," *Trends in Amplification*, vol. 17, no. 3, pp. 189-196, 2013.

[45] Hu, Y. and Kokkinakis, K., "Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners, *Journal of the Acoustical Society of America*, vol. 135, pp. EL22-EL28, 2014.

[46] IEEE., "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[47] Ito, J., Nakatake, M. and Fujita, S., "Hearing ability by telephone of patients with cochlear implants," *Otolaryngology-Head and Neck Surgery*, vol. 121, pp. 802–804, 1999.

[48] ITU-T. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P.862*, 2000.

[49] Kasturi, K. S. "Signal processing strategies for better melody recognition and improved speech understanding in noise for cochlear implants," PhD dissertation, The University of Texas at Dallas, Dallas, 2006.

[50] Kepler, L. J., Terry, M. and Sweetman, R. H., "Telephone usage in the hearing-impaired population," *Ear Hear*, vol. 13, pp. 311–319, 1992.

[51] Kiefer, J., von Ilberg, C., Reimer, B., Knecht, R., Gall, V., Diller, G., Sturzebecher, E., Pfennigdorff, T. and Spelsberg, A., "Results of cochlear implantation in patients with severe to profound hearing loss – implications for patient selection," *International Journal of Audiology*, vol. 37, no. 6, pp. 382-395, 1998.

[52] Kiefer, J., Pok, M., Adunka, O., Stürzebecher, E., Baumgartner, W., Schmidt, M. and Gstoettner, W., "Acoustic simulations of combined electric and acoustic hearing," *Audiology and Neurotology*, vol. 10, pp. 134-144, 2005.

[53] Kim, G., Lu, Y., Hu, Y. and Loizou, P. C., "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Journal of the Acoustical Society of America*. vol. 126, pp. 1486-1494, 2009.

[54] Kjellberg, A., "Effects of reverberation time on the cognitive load in speech communication: Theoretical considerations," *Noise Health*., vol. 7, pp. 11-21, 2004.

[55] Kokkinakis, K., Hazrati, O. and Loizou, P. C., "A channel-selection criterion for suppressing reverberation in cochlear implants," *Journal of the Acoustical Society of America*, vol. 129, pp. 3221-3232, 2011.

[56] Kokkinakis, K. and Loizou, P. C., "Selective-tap blind dereverberation for two-microphone enhancement of reverberant speech," *IEEE Signal Processing Letters*, vol. 16, pp. 961-964, 2009.

[57] Kokkinakis, K. and Loizou, P. C., "The impact of reverberant self-masking and overlap-masking effects on speech intelligibility by cochlear implant listeners," *Journal of the Acoustical Society of America*, vol. 130, pp. 1099–1102, 2011.

[58] Kokkinakis, K., Runge, C., Tahmina, Q. and Hu, Y., "Evaluation of a spectral subtraction strategy to suppress reverberant energy in cochlear implant devices," *Journal of the Acoustical Society of America*, vol. 138, no. 1, pp. 115–124, 2015.

[59] Kong, Y., Stickney, G. S. and Zeng, F., "Speech and melody recognition in binaurally combined acoustic and electric stimulation," *Journal of the Acoustical Society of America*, vol. 117, pp. 1351–1361, 2005.

[60] Kuttruff, H., "*Room Acoustics*" (Taylor & Francis, New York), 2000.

[61] Lebart, K. and Boucher, J. M., "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, pp. 359-66, 2001.

[62] Li, N. and Loizou, P. C., "A glimpsing account for the benefits of simulated combined acoustic and electric hearing," *Journal of the Acoustical Society of America*, vol. 123, pp. 2287–2294, 2008.

[63] Liu, C., Fu, Q. J. and Narayanan, S. S., "Effect of bandwidth extension to telephone speech recognition in cochlear implant users," *Journal of the Acoustical Society of America*, vol. 125, EL77–EL83, 2009.

[64] Liu, C., Azimi, B., Tahmina, Q. and Hu. Y., "Effects of low harmonics on tone identification in natural and vocoded speech", *Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. EL378-EL384, 2012.

[65] Loizou, P. C., "Mimicking the Human Ear," *IEEE Signal Processing Magazine*, vol. 15, no. 5, pp. 101-130, ISSN 1053-5888, September 1998.

[66] Loizou, P.C., "Introduction to Cochlear Implant," *IEEE Engineering in Medical and Biology Magazine*, vol. 18, no. 1, pp. 32-42, ISSN 0739-5175, January/February 1999.

[67] Loizou, P.C., Arthur Lobo and Yi Hu., "Subspace algorithms for noise reduction in cochlear implants (L)," *Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 2791-2793, 2005.

[68] Loizou, P., "Speech enhancement: Theory and practice," *Boca Raton, FL: CRC Press, Taylor Francis Group*, 2007.

[69] Ma. J., Hu,Y. and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band importance functions," *J. Acoustical Society America*, vol. 125, no. 5, pp. 3387–3405, 2009.

[70] Mason, M. and Kokkinakis, K., "Perception of consonants in reverberation and noise by adults fitted with bimodal devices," *Journal of Speech, Language and Hearing Research*, vol. 57, no. 4, pp. 799-899, 2014.

[71] Milchard, A. J. and Cullington, H. E., "An investigation into the effect of limiting the frequency bandwidth of speech on speech recognition in adult cochlear implant users," *International Journal of Audiology*, vol. 43, pp. 356–362, 2004.

[72] Mines, M., Hanson, B. and Shoup, J., "Frequency of occurrence of phonemes in conversational English," Lang Speech **21**, 221–241, 1978.

[73] Miyoshi, M. and Kaneda, Y., "Inverse Filtering of Room Acoustics," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 2, February 1988.

[74] Nabelek, A. K. and Picket, J. M., "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *Journal of Speech, Language and Hearing Research*, vol. 17, pp. 724-739, 1974.

[75] Nabelek, A. K. and Letowski, T. R., "Vowel confusions of hearing impaired listeners under reverberant and non-reverberant conditions," *Speech and Hearing Disorders*, vol. 50, pp. 126-131, 1985.

[76] Nabelek, A. K. and Letowski, T. R., "Similarities of vowels in non-reverberant and reverberant fields," *Journal of the Acoustical Society of America*, vol. 83, pp. 1891-1899, 1988.

[77] Nabelek, A. K., Letowski, T. R. and Tucker, F. M., "Reverberant overlap- and self-masking in consonant identification," *Journal of the Acoustical Society of America*, vol. 86, pp. 1259-1265, 1989.

[78] Neuman, A. C., Wroblewski, M., Hajicek, J. and Rubinstein, A., "Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults," *Ear Hear*, vol. 31, pp. 336-344, 2010.

[79] Nilsson, M., Soli S. and Sullivan J., "Development of Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *Journal of the Acoustical Society of America*, vol. 95, pp. 1085-1099, 1994.

[80] Orabi, A. A., Mawman, D., Al-Zoubi, F., Saeed, S. R., and Ramsden, R. T., "Cochlear implant outcomes and quality of life in the elderly: Manchester experience over 13 years," *Clinical Otolaryngology*, vol. 31(2), pp. 116-122, 2005.

[81] Peter, Jax. and Peter Vary.," On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, pp. 1707–1719, 2003.

[82] Peterson, F. E. and Lehiste, I., "Revised CNC lists for auditory tests," *Journal of Speech and Hearing Disorders*, vol. 27, pp. 62–70, 1962.

[83] Plapous, C., Marro, C. and Scalart, P., "Improving Signal-to-Noise ratio estimation for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 2098-2108, 2006.

[84] Plomp, R. and Mimpen, A.M., "Speech reception threshold for sentences as a function of age and noise level," *Journal of the Acoustical Society of America*, vol. 66, no. 5, pp. 1333-1342, November 1979.

[85] Poissant, S. F., Whitmal, N. A. III., and Freyman, R. L., "Effects of reverberation and masking on speech intelligibility in cochlear implant simulations," *Journal of the Acoustical Society of America*, vol. 119, pp. 1606-1615, 2006.

[86] Polack, J. D., La transmission de l'energie sonore dans les salles," PhD dissertation, University e du Maine, La Mans, France, 1988.

[87] Qin, M. K. and Oxenham, A. J., "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," *Journal of the Acoustical Society of America*, vol. 114, no. 1, pp. 446-454, 2003.

[88] Qin, M. K. and Oxenham, A. J., "Effects of introducing unprocessed low-frequency information on the reception of envelope-vocoder processed speech," *Journal of the Acoustical Society of America*, vol. 119, pp. 2417–2426, 2006.

[89] Rychtarikova, M., Van den Bogaert, T., Vermeir, G. and Wouters, J., "Binaural sound source localization in real and virtual rooms," *Journal of the Audio Engineering Society*, vol. 57, pp. 205-220, 2009.

[90] Spahr, A. J. and Dorman, M. F., "Effects of minimum stimulation settings for the Med EL Tempo + speech processor on speech understanding," *Ear and Hearing*, vol. 26, pp. 2S–6S, 2005.

[91] Spitzer, S., Liss, J., Spahr, A. J., Dorman, M. F. and Lansford, K., "The use of fundamental frequency for lexical segmentation in listeners with cochlear implants," *Journal of the Acoustical Society of America*, vol. 125, EL236–EL241, 2009.

[92] Stevens, K. N., "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," *Journal of the Acoustical Society of America*, vol. 111, pp. 1872-1891, 2009.

[93] Steeneken, H. and Houtgast, T., "A physical method for measuring speech-transmission quality," *J. Acoustical Society America*, vol. 67, no. 1, pp. 318–326, 1980.

[94] Studebaker, G. A., "A rationalized arcsine transform," *Journal of Speech and Hearing Research*, vol. 28, pp. 455–462, 1985.

[95] Stylianou, Y., "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transaction on Speech Audio Process*. vol. 9, pp. 21–29, 2001.

[96] Taal, C. H., Hendriks, R. C., Heusdens, R. and Jensen, J., "A Short-time intelligibility measure for time-frequency weighted noisy speech," *IEEE International Conference on Acoustic Speech and Signal Processing,* March 2010.

[97] Tahmina, Q., Hu, Y. and Chen, F., "Combining harmonic regeneration with noise suppression to improve speech recognition in noise by Cochlear implant listeners", *In Proceedings of the 2013 Conference on Implantable Auditory Prostheses,* 2013.

[98] Tahmina, Q., Hu, Y., Runge, C. and Friedland, D., "Improving speech perception in noise for cochlear implant listeners by combining harmonic regeneration with noise suppression," *In Proceedings of the Association for Research in Otolaryngology 37th Annual Midwinter Meeting*, 2014.

[99] Tahmina, Q., Hu, Y., Runge, C. and Friedland, D., "The contributions of harmonics to the benefits of electroacoustic stimulation", *In Proceedings of the Conference on Implantable Auditory Prostheses*, 2011.

[100] Taylor, B., "Speech-in-noise tests: Including them in your basic test battery," *The Hearing Journal*, vol. 56, no. 1, pp. 40-47, 2003

[101] Terry, M., Bright, K., Durian, M., Kepler, L., Sweetman, R. and Grim, M., "Processing the telephone speech signals for the hearing impaired," *Ear and Hearing*, vol. 13, no. 2, 1992.

[102] Theunissen, M., Swanepoel, D. W. and Hanekom, J., "Sentence recognition in noise: Variables in compilation and interpretation of tests," *International Journal of Audiology*, vol. 8, no. 11, pp. 743-757, 2009.

[103] Turner, C. W., Gantz, B. J., Vidal, C., Behrens, A. and Henry, B. A., "Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing," *Journal of the Acoustical Society of America*, vol. 115, pp. 1729–1735, 2004.

[104] Tyler, R. S., Parkinson, A. J., Woodworth, G. G., Lowder, M. W. and Gantz, B. J., "Performance over time of adult patients using the Ineraid or Nucleus cochlear implant," *Journal of the Acoustical Society*, vol. 102, no. 1, pp. 508-522, 1997.

[105] Vandali, A. E., Whitford, L. A., Plant, K. L. and Clark, G. M., "Speech perception as a function of electrical stimulation rate: Using the Nucleus 24 cochlear implant system," *Ear Hear*, vol. 21, pp. 608-624, 2000.

[106] Van Hoesel, R. and Clark, G., "Evaluation of a portable two-microphone adaptive beamforming speech processor with cochlear implant patients," *Journal of the Acoustical Society of America*, vol. 97, no. 4, pp. 2498-2503, 1995.

[107] von Ilberg, C., Kiefer, J., Tillein, J., Pfenningdorff, T., Hartman, R., Sturzebecher, E. and Klinke, R., "Electric-acoustic stimulation of the auditory system," *ORL, Journal for Otorhinolaryngology and its related specialties, Head and Neck Surgery*, vol. 61, pp. 334–340, 1999.

[108] Whitmal, N. A. and Poissant, S. F., "Effects of source-to-listener distance and masking on perception of cochlear implant processed speech in reverberant rooms," *Journal of the Acoustical Society of America*, vol. 126, pp. 2556-2569, 2009.

[109] Wightman, Frederic L. and Kistler, Doris J., "Headphone simulation of free-field listening. I: Stimulus synthesis," *Journal of the Acoustical Society of America*, vol. 85, no. 2, February 1989.

[110] Wouters, J. and Vanden Berghe J., "Speech recognition in noise for cochlear implantees with a two-microphone monaural adaptive noise reduction system," *Ear Hear*, vol.22, pp. 420-430, 2001.

[111] Yang, L. and Fu, Q., "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *Journal of the Acoustical Society of America, (L),* vol. 117, no. 3, pp. 1001-1004, 2005.

[112] Zavarehei, E., Vaseghi, S. and Yan, Q., "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," *IEEE Transaction on Audio, Speech, Language Processing*, vol. 15, pp. 1194–1203, 2007.

[113] Zhang, T., Dorman, M. F. and Spahr, A. J., "Frequency overlap between electric and acoustic stimulation and speech-perception benefit in patients with combined electric and acoustic stimulation," *Ear and Hearing*, vol. 31, pp. 195–201, 2010a.

[114] Zhang, T., Dorman, M. F. and Spahr, A. J., "Information from the voice fundamental frequency (F0) region accounts for the majority of the benefit when acoustic stimulation is added to electric stimulation," *Ear and Hearing*, vol. 31, pp. 63-69, 2010b.

[115] Zwicker, E., Feldtkeller, R., Das Ohr als Nachrichtenempfänger, *S. Hirzel Verlag*, Stuttgart, 1967.

APPENDICES

APPENDIX A

ITU–T IRS Filter

ITU-T is a specialized sector of International Telecommunication Union (ITU). It is responsible for making recommendations that describes methods and procedures on subjective and objective performance evaluation of digital speech codes and telecommunication networks [48]. Subjective evaluation of telephone networks might be conducted using subjective listening tests or conversational methods. Usually, for practical reasons, listening tests are the most feasible for subjective testing during the development of speech coding algorithms.

Researchers rely on the subjective and objective listening tests for evaluation of speech coding algorithms when real-time implementation of speech codecs is not available. In order to evaluate the speech quality, the receiving frequency characteristics of telephone handsets needs to be simulated. In order to filter the speech signals, a modified Intermediate Reference System (IRS) filters from ITU-T P.862 is used in this study. To estimate the subjective quality from the listening tests which use listening equipment that conforms to the IRS or modified IRS receive characteristics, the objective measurement techniques are described by the ITU-T P.862 recommendation from International Telecommunication Union. IRS filtering is used to model the signals that are actually heard by the human subjects. IRS filter is used in this dissertation to filter the band-limited telephone speech and its frequency response is shown in Figure A.1.

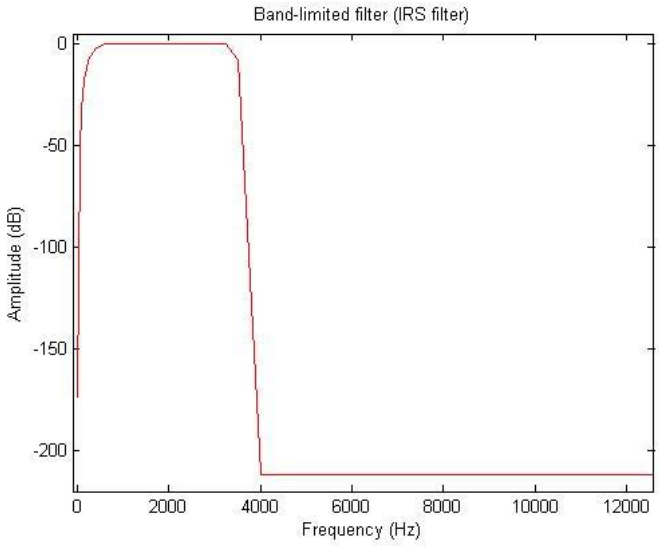


Figure A.1. Band-limited IRS filter used in this study.

To implement the IRS filtering, the time domain speech signal is transformed into the frequency domain using FFT over the length of the speech file. The gain is then computed by filtering the signal in frequency domain using piecewise linear response and interpolation at each frequency. Then, inverse FFT is performed over the length of the speech signal to obtain the filtered signal. The filtered signals are then normalized over the speech signals to be used for the listening tests.

APPENDIX B

Head Related Transfer functions

Human beings perceive sounds in different ways depending on the size and shape of the head, pinna and torso. There are environmental factors that alter the sounds received by the listeners and therefore are modeled by the head related impulse response. The ratio of amplitude of received to incident sounds for a specific angle of incidence with respect to the listener is referred to as a head related transfer function. In acoustical reverberant environments, the sound source, listener and the room impulse response combine to form an audible stimuli. The method to simulate reverberant acoustical condition is used from the study by Rychtarikova *et al.* [89].

HRTF measurements:

Generally, Head related transfer functions are measured by placing mannequins at a fixed distance from the listener's head for both left and right ears. Transfer functions are measured at several different horizontal azimuths and vertical elevations in degrees or radians. In order to perform system identification, a known stimulus is played through the loud speakers placed at a distance from the subject's head [109] and at a specified azimuth θ, elevation φ. Also, probe tube microphones are partially inserted into a subject's ears to measure the empirical Head-related transfer function [11]. Speaker and microphone transfer functions are removed from the measured transfer function. Common transfer functions are those that are measured the same at all locations and are thus removed from the raw measurements. The directional transfer function (DTF) at azimuth θ and elevation φ is referred to as HRTF and contains spectral cues responsible for spatial

hearing. Mathematically, the directional transfer function from the stimulus signal and common transfer function is computed as follows [11]. For a known stimulus signal $s(t)$ at azimuth θ and elevation φ and $c(t)$ a known CTF, if $dtf_{l,\theta,\emptyset}\,(t)$and $dtf_{r,\theta,\emptyset}\,(t)$ are the unknown left and right ear DTFs, then the signals recorded from left and right microphones $m_{l,\theta,\emptyset}\,(t)$ and $m_{r,\theta,\emptyset}\,(t)$ can be defined;

$$m_{l,\theta,\emptyset}\,(t) = s(t) * c(t) * dtf_{l,\theta,\emptyset}\,(t)\;;\quad m_{r,\theta,\emptyset}\,(t) = s(t) * c(t) * dtf_{r,\theta,\emptyset}\,(t) \tag{B.1}$$

In frequency domain;

$$M_{l,\theta,\emptyset}\,(\omega) = S(\omega)\,C(\omega)\,DTF_{l,\theta,\emptyset}\,(\omega)\;;\quad M_{r,\theta,\emptyset}\,(\omega) = S(\omega)\,C(\omega)\;\,DTF_{r,\theta,\emptyset}\,(\omega) \tag{B.2}$$

Here, $c(t)$ is assumed to be invariant in the spatial domain and therefore could be computed from the recording apparatus and spectrally averaged values of $m_{l,\theta,\emptyset}\,(t)$ and $m_{r,\theta,\emptyset}\,(t)$ for several locations. Therefore, the left and right directional transfer functions can be computed as:

$$\left|DTF_{l,\theta,\emptyset}(\omega)\right| = \frac{\left|M_{l,\theta,\emptyset}\,(\omega)\right|}{|S(\omega)\,||C(\omega)\,|}\;;\quad \left|DTF_{r,\theta,\emptyset}(\omega)\right| = \frac{\left|M_{r\,\theta,\emptyset}\,(\omega)\right|}{|S(\omega)\,||C(\omega)\,|} \tag{B.3}$$

And,

$$< DTF_{l,\theta,\emptyset}\,(\omega) = < M_{l,\theta,\emptyset}(\omega) - < S(\omega) - < C(\omega)\;;$$

$$< DTF_{r,\theta,\emptyset}\,(\omega) = < M_{r,\theta,\emptyset}(\omega) - < S(\omega) - < C(\omega) \tag{B.4}$$

Finally, using (B.3) and (B.4), we computer the DTFs as:

$$DTF_{l,\theta,\emptyset}(\omega) = \left|DTF_{l,\theta,\emptyset}(\omega)\right| exp(j < DTF_{l,\theta,\emptyset}\,(\omega))\;;$$

$$DTF_{r,\theta,\emptyset}(\omega) = \left|DTF_{r,\theta,\emptyset}(\omega)\right| exp(j < DTF_{r,\theta,\emptyset}\,(\omega)) \tag{B.5}$$

By performing the inverse FFT of the (B.5), the directional transfer functions was computed as follows:

$$dtf_{l,\theta,\emptyset}\,(t) = F^{-1}(DTF_{l,\theta,\emptyset}(\omega))_{l,\theta,\emptyset}\,(\omega));$$

$$dtf_{r,\theta,\emptyset}\,(t) = F^{-1}(DTF_{r,\theta,\emptyset}(\omega))_{l,\theta,\emptyset}\,(\omega)) \quad \text{(B.6)}$$

HRTF measurements are usually obtained from a CORTEX® MK2 manikin artificial head placed in the middle of the ring of 2.0m inner diameter in an anechoic room. Thirteen single-cone loudspeakers (FOSTEX® 6301B) of 10-cm diameter were placed every 15^{o} in the frontal plane. By transmitting a logarithmic frequency sweep using a 2-channel sound card (VX POCKET 440 DIGIGRAM®) and DIRAC 3.1 software type 7841 (Bruel and Kjaer Sound and Vibration Measurement Systems), the impulse response for every angle for left and right year are computed. Thereafter the recorded sounds are normalized by the transmitted sweep spectrum obtained from the hardware loop-back calibration process. Finally, the HRTF for both left and right ears are computed using an inverse Fourier transform of the spectrum.

APPENDIX C

OBJECTIVE PERFORMANCE MEASURES

Objective measures are efficient, unbiased, time-saving and cost-effective methods to measure the quality of speech. These methods are classified into two categories as intrusive and non-intrusive measures based on whether the reference signal is provided or not. Intrusive methods compare the reference speech signals with the degraded speech signals to provide an estimation of the speech quality perceived by the subjects. Non-intrusive methods use predictions to compare the reference signal and the degraded signal. Several intrusive measures have been developed over the years both for the purpose of quality and intelligibility prediction. In this appendix, we will describe two objective measures used in this dissertation.

C.1 PESQ (Perceptual Evaluation of Speech Quality)

Perceptual Evaluation of Speech Quality (PESQ) is an International Telecommunications Union (ITU-T) P.862 Recommendation for speech quality assessment of narrow-band and wide-band speech which is used for speech codecs and also end-to-end measurements [48]. In this appendix, we will provide an overview of the model for PESQ algorithm along with the signal processing involved in the measurement of the speech quality. Here, in this section we will be discussing the model described in PESQ algorithm [48].

The basic idea of the PESQ algorithm is to compare degraded speech signal with that of reference speech signal and provide a speech-quality score. PESQ measure predicts the perceived quality obtained from a subjective listening test. In order to compare the degraded signal with the original signal, the algorithm transforms both the signals into a psychophysical representation of audio signals in the human auditory system. Time and level alignment, time- frequency mapping,

frequency warping, and compressive loudness scaling are used to obtain these representations as described in [48]. In the cognitive model, complex non-linear calculations are performed to compute the speech quality degradation represented by the disturbance metric between the psychophysical representations of the reference and degraded speech signals. Two error parameters from the cognitive model are then combined to produce an objective listening quality measure called a mean opinion scores (MOS). Figure C.1 shows an overview of the PESQ model.

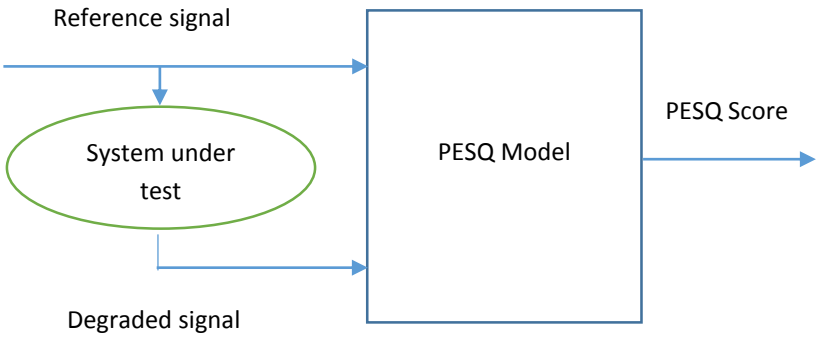


Figure C.1. Block diagram of PESQ method used in this study for objective measures.

PESQ algorithm:

PESQ score for the reference and degraded signal can be found using the perceptual model. In this dissertation, PESQ is used to assess the quality of speech in the presence of background noise. In order to consider the subjective disturbance in the context of absolute category rating (ACR), the original signal should be clean, and but the signal under test should be corrupted by noise. PESQ is used to validate the improvement of the speech enhancement methods for the corrupted and noisy signals.

Computation of the PESQ score:

- Level and time alignment: Initially, the reference signal, *X(t)* and degraded signal, *Y(t)* are scaled to maintain constant power level. Different gains are computed from the filtered samples and are applied to reference and degraded signals. To implement the perceptual

model of the human evaluation for speech quality, an IRS receive characteristics are used. Time alignment is performed using envelope-based delay estimation so that degrade speech match the reference speech.

- Auditory Transform: In PESQ, a psychoacoustic model is implemented using time-frequency mapping, frequency warping and comprehensive loudness scaling. Using a Hanning window of 32 ms, a short-time FFT of the signal is computed. The power spectrum of the complex FFT components are computed for the original and degraded signals: $PX_{WIRSS}(f)_n$and $PX_{WIRSS}(f)_n$. After normalization and filtering, frequency warping function is used to map the frequency scale in Hertz to pitch scale in Bark to compute the pitch power densities $PX_{WIRSS}(f)_n$ and $PY_{WIRSS}(f)_n$. In each frame, a sum of all *n* pitch power densities that exceeds the absolute hearing threshold is determined. Distorted pitch power density is multiplied by the ratio of original to degraded power to compute the gain compensated pitch power density for each frame. Original and degraded pitch power densities are transformed to a Sone loudness scale using Zwicker's law [115] to obtain the loudness densities $LX(f)_n$a and $LY(f)_n$.

$$LX(f)_n = S_l.\left(\frac{P_0(f)}{0.5}\right)^{\gamma}\left[\left(0.5 + 0.5\,\frac{PPX'_{WIRSS}(f)_n}{P_0(f)}\right)^{\gamma} - 1\right] \qquad \text{(C.1.1)}$$

$$LY(f)_n = S_l.\left(\frac{P_0(f)}{0.5}\right)^{\gamma}\left[\left(0.5 + 0.5\,\frac{PPY'_{WIRSS}(f)_n}{P_0(f)}\right)^{\gamma} - 1\right] \qquad \text{(C.1.2)}$$

Where $P_0(f)$ is the absolute threshold and S_l is the loudness scaling factor. Above 4 Bark, the Zwicker power, γ, is 0.23, and below 4 Bark, the recruitment effect is observed if the Zwicker power is increased.

- Disturbance processing: The difference between the original and the distorted loudness densities is computed for each frame and the difference array is called the raw disturbance density. Disturbance density $D(f)_n$and asymmetrical disturbance density $DA(f)_n$are calculated to model the asymmetric effect caused by the codec in each frame, which are then combined in the frequency domain:

$$D_n = M_n \sqrt[3]{\sum_{f=1,\ldots\ldots No.of\ Bark\ Bands} \left(|D(f)_n| W_f\right)^3} \qquad \text{(C.1.3)}$$

Where *Mn* is a multiplication factor, 1/(power of original frame plus a constant)0.04. These combined values, D_n and DA_n, are called frame disturbances. If the decrease in the delay is larger than half a window (16 ms), the frame disturbances are ignored. The resulting frame disturbances are D'_n and DA'_n.

$$DA_n = M_n \sum_{f=1,\ldots\ldots No.of\ Bark\ Bands} (|DA(f)_n| W_f) \qquad \text{(C.1.4)}$$

If the frame disturbance of consecutive frames are above a threshold, then those are considered as bad intervals. By maximizing the cross-correlation between the absolute original signal and absolute degraded signal adjusted depending on the pre-processing, a new delay value is estimated. The processing continues until the maximal cross-correlation falls below a threshold, which means that the interval is matching noise against noise. The result is the final frame disturbances D''_n (measure of absolute audible error) and DA''_n (measure of errors for frames louder than the reference) that are used to calculate the perceived quality. The frame disturbance and asymmetrical frame disturbance values are then combined over the active interval of the speech file to determine the average values.

Thus, PESQ score is a linear combination of average of frame disturbance and average of asymmetrical frame disturbance.

The PESQ score is a listening quality with mean opinion score between 1.0 and 4.5, which is similar to the normal range found in an ACR experiments.

C.2 Short-time objective intelligibility measure (STOI)

Taal, C. H [96] designed a model uses TF decomposition based on DFT to measure the intermediate intelligibility for short-time TF-region. Here we are discussing this model for the measure of speech intelligibility described well in [96]. The model assumed the clean and processed signals are time aligned with a sample-rate of 10000 Hz. The model first computes an FFT by applying a Hanning-window of length 256 samples with 50% overlap and to obtain the TF-representation of the signal. Then, DFT-bins are grouped to perform one-third octave band analysis. If *x* and *y* represent the clean and the processed signals, then for *m*th frame of the clean signal, $\hat{x}\ (k, m)$ denote the *k*th DFT-bin. The TF-unit for the *j*th one-third octave band, is denoted as:

$$X_j\ (m) = \sqrt{\sum_{k=k1(j)}^{k2(j)-1} |\hat{x}(k,m)|^2} \tag{C.2.1}$$

where *k*1 and *k*2 denote the one-third octave band edges rounded to the nearest DFT-bin.

Similarly, the TF-unit of the processed speech is obtained as:

$$Y_j\ (m) = \sqrt{\sum_{k=k1(j)}^{k2(j)-1} |\hat{y}(k,m)|^2} \tag{C.2.2}$$

These TF units of the processed signal $Y_j\ (m)$ is then normalized by scaling a signals with a factor $\alpha = \sqrt{\frac{\sum_n X_j(n)^2}{\sum_n Y_j(n)^2}}$ such that the energy equals the clean energy for that specific TF unit.

Normalization process includes clipping of the $\alpha Y_j\ (n)$ in order to bound the signal-to-distortion ratio, which is defined as:

$$SDR_j\ (n) = \ 10\ log_{10}(\frac{X_j\ (n)^2}{(\ \alpha Y_j\ (n) - X_j\ (n))^2}) \tag{C.2.3}$$

So, the normalized and clipped TF unit can be represented as:

$$Y' = \max\big(\min\big(\alpha Y,\ X + 10^{-\beta/20}\ X\big), X - 10^{-\beta/20}\ X\big) \tag{C.2.4}$$

Where β denotes the lower SNR bound. An estimate of a linear correlation coefficient between the clean and processed TF-units is called the intermediate intelligibility measure,

$$d_j\ (m)\ \ = \ \ \frac{\sum_n\big(X_j(n)-\frac{1}{N}\sum_l X_j(l)\big)\ \big(Y_j'(n)-\frac{1}{N}\sum_l Y_j'(l)\big)}{\sqrt{\sum_n\big(X_j(n)-\frac{1}{N}\sum_l X_j(l)\big)^2\ \sum_n\big(Y_j'(n)-\frac{1}{N}\sum_l Y_j'(l)\big)^2}} \tag{C.2.5}$$

where $l \in M$. For one TF unit, $dj\ (m)$, depends on a N consecutive TF-units from both $Xj(n)$ and $Yj(n)$, where $n \in M$ and $M=\{(m-N+1)\ ,\ (m-N+2)\ ,\ ...,m-1,m\}$. An average of the intermediate intelligibility measures over all bands and frames gives the objective intelligibility measure.

$$d = \ \frac{1}{JM}\ \sum_{j,m} d_j(m) \tag{C.2.6}$$

where M is the total number of frames and J is the number of one-third octave bands. Intermediate measure depends on speech information from the last ≈400 ms since maximum correlation is obtained with β=-15 and N=30.

Curriculum Vitae

Qudsia Tahmina

Place of birth: Hyderabad, India

Education:

PhD, Electrical Engineering
University of Wisconsin-Milwaukee, Milwaukee, WI
December, 2015

M.S., Electrical and Computer Engineering
Purdue University Calumet, Hammond, IN
December, 2007

B.S., Information Technology
Osmania University, Hyderabad, India
May, 2005

Dissertation Title: Coding Strategies for Cochlear Implants under Adverse Environments

Research Experience

Research Assistant, University of Wisconsin-Milwaukee (September 2009-December 2014)

Teaching Experience

Faculty of Engineering, Grantham University (January 2015 - Present)
Teaching Assistant, University of Wisconsin-Milwaukee (January 2010 - August 2014)
Teaching Assistant, Purdue University Calumet (August 2005 - December 2007)

Awards and Fellowships

Travel Award recipient sponsored by the Association of Research in Otolaryngology and University of Wisconsin-Milwaukee (Feb 2014)
Nominated for the Distinguished Graduate Student Fellowship (DGSF) award at University of Wisconsin-Milwaukee (2013-2014)
Chancellor's Graduate Student Award recipient at University of Wisconsin-Milwaukee (2009-2013)

Publications

Kokkinakis, K., Runge, C., Tahmina, Q. and Hu, Y., "Evaluation of a spectral subtraction strategy to suppress reverberant energy in cochlear implant devices," *Journal of the Acoustical Society of America*, vol. 138, no. 1, pp. 115–124, 2015.

Hu, Y., Tahmina, Q., Runge, C. and Friedland, D. R., " The perception of telephone-processed speech by combined electric and acoustic stimulation," *Trends in Amplification*, vol. 17, no. 3, pp. 189-196, 2013.

Liu, C., Azimi, B., Tahmina, Q. and Hu, Y., "Effects of low harmonics on tone identification in natural and vocoded speech," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. EL378-EL384, 2012.

Chen, F., Wong, L. L., Tahmina, Q., Azimi, B. and Hu, Y., "The effects of binaural spectral resolution mismatch on Mandarin speech perception in simulated electric hearing," *The Journal of the Acoustical Society of America*, 132, no. 2, pp. EL142- EL148, 2012.

Tahmina, Q., Chen, F. and Hu Yi, "Perceptual contribution of vowels and consonants to sentence intelligibility by cochlear implant users," *Invited Lecture Presentation for the International Symposium on Integrated Circuits*, 2014.

Tahmina, Q., Hu, H., Runge, C. and Friedland, D. R.,"Improving speech perception in noise for Cochlear Implant listeners by combining Harmonic regeneration with noise suppression," *In Proceedings of the Association for Research in Otolaryngology*, 2014.

Tahmina, Q., Hu, Y. and Chen, F., "Combining harmonic regeneration with noise suppression to improve speech recognition in noise by Cochlear implant listeners," *In Proceedings of the Conference on Implantable Auditory Prostheses*, 2013.

Tahmina, Q., Bhandary, M., Azimi, B., Hu, Y., Utianski, R. L. and Liss, J., "The effects of visual information on speech perception in noise by electroacoustic hearing," *The Journal of the Acoustical Society of America*, vol. 132, 2012.

Tahmina, Q., Hu, Y., Runge, C. and Friedland, D. R, "The contributions of harmonics to the benefits of electroacoustic stimulation," *In Proceedings of the 2011 Conference on Implantable Auditory Prostheses*, 2011.