Implementation and Evaluation of Spectral Subtraction with Minimum Statistics using WOLA and FFT Modulated Filter Banks

This thesis is presented as part of Degree of
Master of Science in Electrical Engineering with emphasis on Signal Processing

Peddi Srinivas Rao
Vallabhaneni Sreelatha

Blekinge Institute of Technology
January 2014

Blekinge Institute of Technology
School of Engineering (ING)
Department of Signal Processing
Supervisor: Dr. NedelkoGrbic
Examiner: Dr. Sven Johansson.

This thesis is submitted to the School of Engineering at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering with emphasis on Signal Processing.

Contact Information:

Authors:

Peddi Srinivas Rao
E-mail:srpe10@student.bth.se

Vallabhaneni Sreelatha
E-mail: srva10@student.bth.se

Supervisor:
Dr. NedelkoGrbić
ING School of Engineering

Examiner:
Dr. Sven Johansson
ING School of Engineering

School of Engineering
Blekinge Institute of Technology
371 79 Karlskrona
Sweden

Internet : www.bth.se/ing
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

ABSTRACT

In communication system environment speech signal is corrupted due to presence of additive acoustic noise, so with this distortion the effective communication is degraded in terms of the quality and intelligibility of speech. Now present research is going how effectively acoustic noise can be eliminated without affecting the original speech quality, this tends to be our challenging in this current research thesis work. Here this work proposes multi-tiered detection method that is based on time-frequency analysis (i.e. filter banks concept) of the noisy speech signals, by using standard speech enhancement method based on the proven spectral subtraction, for single channel speech data and for a wide range of noise types at various noise levels.

There were various variants have been introduced to standard spectral subtraction proposed by S.F.Boll. In this thesis we designed and implemented a novel approach of Spectral Subtraction based on Minimum Statistics [MinSSS]. This means that the power spectrum of the non-stationary noise signal is estimated by finding the minimum values of a smoothed power spectrum of the noisy speech signal and thus circumvents the speech activity detection problem. This approach is also capable of dealing with non-stationary noise signals. In order to analyze the system in time frequency domain, we have implemented two different filter bank approaches such as Weighted OverLap Added (WOLA) and Fast Fourier Transform Modulated (FFTMod).

The proposed systems were implemented and evaluated offline using simulation tool Matlab and then validated their performances based on the objective quality measures such as Signal to Noise Ratio Improvement (SNRI) and Perceptual Evaluation Speech Quality (PESQ) measure. The systems were tested with a pure speech combination of male and female sampled at 8 kHz, these signals were corrupted with various kinds of noises at different noise power levels. The MinSSS algorithm implemented using FFTMod filter bank approach outperforms when compared the WOLA filter bank approach.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude and thanks to thesis supervisor Dr.Nedelko Grbic for giving us a wonderful opportunity to do thesis research work in speech processing filed under his guidance. We also thank for his continuous feedback and encouragement throughout the thesis work, without this it would have been difficult in doing this gigantic research work successfully.

We extend our sincere appreciation and thanks to our fellow students at Blekinge Institute of Technology, Sweden for their continuous and suggestions in solving different issues while doing this thesis work.

Finally, we would like to extend immense gratitude and wholehearted thanks to our parents for their moral support and encouragement throughout my career. We also take opportunity to thank our siblings, and our mutual friends for their support and encouragement in completing this thesis work

Contents

Abstract 3

Acknowledgements 4

List of figures 7

List of tables 8

List of acronyms and abbreviations 9

CHAPTER-1 10

I. INTRODCTION 10

1.1 Overview 10

1.2 Overview of the proposed system 11

1.3 Thesis Organization 12

2. BACKGROUND THEORIES 13

2.1 Introduction 13

2.2 Fundamentals of speech production 13

2.3 Filter bank: 16

2.3.1 Digital filter bank: 16

2.4 Speech Enhancement 18

2.4.1 Single Channel Speech Enhancement 18

2.5 Spectral Subtraction 21

2.5.1. Principle of Spectral Subtraction 21

2.5.2 Drawbacks of Spectral Subtraction 23

3. DESIGN, IMPLEMENTATION and EVALUATION OF FILTER BANKS 25

3.1 Design of WOLA filter bank: 25

3.1.1 An analysis bank based on DFT: 25

3.1.2 A synthesis bank based on inverse DFT: 26

3.1.3. Implementation and Evaluation of Filter Bank 27

3.2. Design of FFT-Modulated Filter Bank 28

3.2.2 Implementation and Evaluation of Filter Bank 32

4. DESIGN AND IMPLEMENTATION OF MinSSS 33

4.1. Spectral Subtraction using minimum statistics 33

4.1.1. Description 34

4.1.2 Subband Power Estimation 34

4.1.3 Subband Noise Power Estimation 35

4.1.4 Compute SNR 35
4.1.5 Subtraction Rule 36
4.1.6 Description of parameters 36
4.2 Implementation in Matlab 38
CHAPTER-5 40
5. Simulation results and analysis 40
5.1 Evaluation Parameters 40
5.1.1 Signal-to-Noise-Ratio (SNR) 41
5.1.2 Perceptual Evaluation of Speech Quality (PESQ). 41
5.1.3 Total Noise Level Reduction (TNLR) 41
5.1.4 SNRI-to-NPLR difference (DSN) 42
5.2. WOLA-FB Minimum Statistics Spectral Subtraction 43
5.3 FFT Modulated Filter Bank Minimum Statistics Spectral Subtraction 44
5.4 Comparison of WOLA-FB MinSSS and FFTMod-FB MinSSS 46
CHAPTER-6 53
6. Conclusion and future work 53
6.1 Conclusion 53
6.2 Future work: 53
CHAPTER-7 54
7.Bibliography 54

LIST OF FIGURES

Figure 1: Overview of speech enhancement system designed……11
Figure 2.1: Human speech production system……14
Figure 2.2 Phonemes in American English……15
Figure 2.3 Analysis-Synthesis filter banks……..16
Figure 2.4 Generalized Filter Bank structure……..17
Figure 2.5 Short-time spectral magnitude enhancement system……21
Figure 2.6 General form of spectral Subtraction……22
Figure 3.1 Analysis stage of WOLA filter……..26
Figure 3.2 Synthesis stage of WOLA filter……27
Figure 3.3 Original and Reconstructed Speech signal of WOLA-FB……28
Figure 3.4 Uniform FFT Modulated Analysis FB……...31
Figure 3.5 Uniform FFT modulated Synthesis FB stage……31
Figure 3.6 Original and Reconstructed Speech signal using FFT Modulated FB……32
Figure 4.1 Block diagram of Spectral subtraction by minimum statistics……... 34
Figure 4.2 Estimate of smoothed power signal and the estimate of noise floor…….. 37
Figure 4.3 Spectral Subtraction using Filter Bank…….. 38
Figure 5.1 SNR-I for different noises at different dB levels using WOLA-FB and MinSSS…… 44
Figure 5.2 SNR-I for different noises at different dB levels using FFT-Modulated FB and MinSSS……..... 46
Figure 5.3 PESQ-I for white noise at different dB levels for WOLA-MinSSS and FFTMod-MinSSS……..... .48
Figure 5.4 SNR-I for white noise at different dB levels for WOLA-MinSSS and FFTMod-MinSSS.48
Figure 5.5 PESQ-I for Pink noise at different dB levels for WOLA-MinSSS andFFTMod-MinSSSS…….....49
Figure 5.6 SNR-I for Pink noise at various dB levels for WOLA-MinSSS and FFTMod-MinSSSS..49
Figure 5.7 PESQ-I for car noise at various dB levels for WOLA-MinSSS and FFTMod-MinSSS….50
Figure 5.8 SNR-I for car noise at various dB levels for WOLA-MinSSS and FFTMod-MinSSS…...50
Figure 5.9 PESQ-I for Engine noise at various dB levels for WOLA-MinSSS and FFTMod-MinSSS…….51
Figure 5.10 SNR-I for Engine noise at various dB levels for WOLA-MinSSS and FFTMod-MinSSS……..51
Figure 5.11 PESQ-I for Factory at various dB levels for WOLA-MinSSS and FFTMod-MinSSS…. 52
Figure 5.12 SNR-I for Factory noise at various dB levels for WOLA-MinSSS and FFTMod-MinSSS……..52

LIST OF TABLES

Table 4.1 Optimum values chosen for MinSSS algorthim…… 39

Table 5.1 MOS (Mean Opinion Score) Scores……. 41

Table 5.2 SNR-I, PESQ-I, TLNR, DSN for all noises at different dB levels using WOLA-FB And MinSSS…… 43

Table 5.3 SNR-I, PESQ-I, TLNR, DSN for all noises at different dB levels using FFT-modulated FB and MinSSS……..45

Table 5.4 Comparison of SNR-i and PESQ-I between WOLA and FFT-Modulated FB's for different noises at various levels……...47

LIST OF ACRONYMS AND ABBREVIATIONS

FB	Filter Bank
PF	PolyphaseFilter
DFT	Discrete Fourier Transform
IDFT	Inverse Discrete Fourier Transform
FFT	Fast Fourier Transform
IFFT	Inverse Fast Fourier Transform
PSD	Power Spectral Density
ARMA	Auto Regressive Moving Average
AR	Auto Regressive
MA	Moving Average
WOLA	Weighted OverLap Added
FFTMod	FFT Modulated
MinSSS	Minimum Statistics Spectral Subtraction
STSA	Short Time Spectral Attenuation
VAD	Voice Activity Detector
osub	over subtraction factor
subf	Spectral Floor constant
SNR-I	Signal to Noise Ration Improvement
PESQ	Perceptual Evaluation of Speech Quality
PESQ-I	Perceptual Evaluation of Speech Quality Improvement
TNLP	Total Noise Level Reduction

CHAPTER-1

I. INTRODCTION

1.1Overview

Today world is trending towards advancing of the speech communication systems at rapid speed, the communication systems such as hand free devices i.e. mobiles, laptops etc. These devices were being used as sound capturing devices in various applications such as of meetings, video conferencing and lectures for archival purposes. These devices often faces a unique source one with randomly occurring high-energy noise bursts embedded into a stationary background noise, resulting into a noise consisting of both stationery and non-stationery noises, whose presence of these kind of noises in speech signal can result in appreciable degradation in both the quality and intelligibility. The background may be noise-like such as in aircraft, street noise, etc., or may be speech-like such as an environment with competing speakers. These kind of background acoustic noises can corrupt voice signals in communication systems, interfering with message interpretation and understanding. This in turn impedes other applications such as speech coding, speech recognition and speaker identification, mobile telephony, air-traffic radio control and voice relay networks. Speech enhancement techniques are concerned with algorithms that mitigate these unwanted noise effects and thus improve signal quality.

Many algorithms have been introduced and developed for single-channel speech enhancement to improve the perceptual quality of the speech signals from the corrupted input signals in communication systems [1, 2]. These algorithms are generally based on short-time spectral attenuation (STSA). The problem of enhancing speech degraded by additive background noise, when only a single channel is available, remains challenging due to the limitations of existing methods' in realistic noise conditions. It is generally difficult to restore desired signal without distorting speech signal and the performance is limited by the trade-off between speech distortion and noise reduction. The most common scenario is the single channel system [3], where noise and speech come from the individual sources and a microphone records both speech and noise. This is the difficult situation to handle because, speech and noise are highly uncorrelated with each other in microphone signal. The computational complexity and cost of implementation in real-time applications such as mobile communications, hearing aids, intelligent hearing protectors is foremost important issues in proposing a speech enhancement algorithm for that

application. The spectral subtraction is one of the ways for speech enhancement [9]. The spectral subtraction algorithm principle involves, in estimating of noise power spectrum from the noisy speech power spectrum during noise/speech-pause regions frame wise in the signal whereas during speech frames it estimates average of previous noise frames power spectra's, finally the clean speech power spectrum is estimated by subtracting the noise power spectrum from the noisy speech power spectrum. Since, last few decades many researchers have been carried out on the spectral subtraction based methods because of its simplicity and ease of implementation on portable devices such as mobile communications [3].

In this thesis, we focus on the reduction of extreme different background noises randomly occurring high-energy such as pink, white, engine, factory babble and music by using Minimum Statistics Spectral Subtraction [MinSSS] based on two different filter bank approaches such as Weighted OverLap Added [WOLA] and FFT Modulated [FFTMod] filter banks.

1.2 Overview of the proposed system

The block diagram of the proposed method is presented in Figure 1. The signal X*(n)* is mixed with additive background noise to pure speech signal sampled at 8 kHz. This noisy signal will undergoes through analysis filter bank stage for time-frequency analysis so that we can analyze in frequency domain, whose frequency bins ($X_0[n], X_2[n], X_2[n] \dots X_{k1}[n]$) are undergo through gain function *G* ($g_0, g_1, g_2 \dots g_{k1}$) of speech enhancement algorithm in order to estimate the pure speech spectrum from noisy signal spectrum, then to reconstruct the estimated speech spectrum in time domain we use synthesis filter, finally we obtains the estimated speech signal in time domain *Y[n]*. Now to measure the performance of our system we will perform various quality analysis measures on estimated speech over the actual speech.

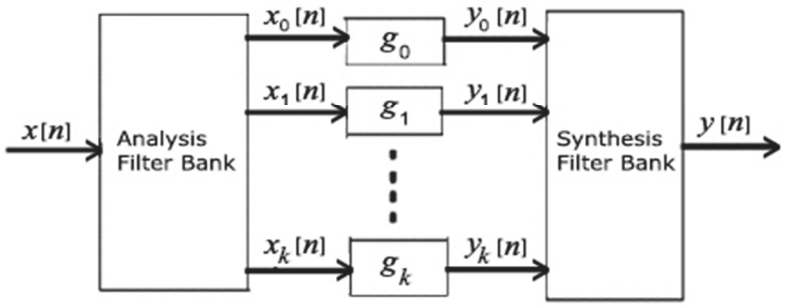

Figure 1: Overview of speech enhancement system designed

1.3 Thesis Organization

The thesis report is divided into five chapters. The remaining paper is organized as follows; Chapter 2 provides information about the speech production model, filter bank concepts and single channel speech enhancement technique i.e. spectral subtraction. In the chapter, we discuss design, implementation of filter banks and then evaluate the designed filter banks. The theory and implementation of the MinSSS algorithm is presented in Chapter 4 and in chapter 5 we evaluate the system implemented i.e. MinSSS using WOLA-FB and FFTMod-FB using four quality parameters. Finally, in chapter 6 the thesis is concluded and provides future research direction on MinSSS algorithm.

CHAPTER 2

2. BACKGROUND THEORIES

2.1 Introduction

In the past decades, research in the field of speech enhancement has focused on the suppression of additive background noise [4, 5, 6]. From the point of view of signal processing, additive noise is easier to deal with than convolutive noise or nonlinear disturbances [7]. The ultimate goal of speech enhancement is to eliminate the additive noise present in speech signal and restore the speech signal to its original form. Several methods have been developed as a result of these research efforts [15]. Most of these methods have been developed with some or the other auditory, perceptual or statistical constraints placed on the speech and noise signals. However, in real world situations, it is very difficult to reliably predict the characteristics of the interfering noise signal or the exact characteristics of the speech waveform. Hence, in effect, the speech enhancements methods are sub-optimal and can only reduce the amount of noise in the signal to some extent. Due to the sub-optimal nature of these methods, some of the speech signal can be distorted during the process. Hence, there is a trade-off between distortions in the processed speech and the amount of noise suppressed. The effectiveness of the speech enhancement system can therefore be measured based on how well it performs in light of this trade-off.

This chapter presents reviews on the production of speech in humans, concept of filter bank and a literature review of the different speech enhancement methods used to date. The family of subtractive type enhancement methods is discussed in more detail.

2.2 Fundamentals of speech production

Speech, a dynamic, information-bearing signal, is also called the acoustic waveform. These waves are produced due to the sound pressure generated in the mouth of the speaker as a result of some sequence of coordinated movements of a series of structures in the human vocal system. The branch of science that deals with the dynamics and production of the human sound is called phonetics [7]. The process of speech communication involves the production of the acoustic wave by the speaker and the perception of the signal by the listener. Though the process of speech perception still largely remains a mystery to the scientific world, the process of speech production has been well researched and understood. A sound knowledge of the processes involved in the production and perception of speech is

necessary for engineers to develop suitable methods to represent and transform the acoustic signals to achieve the desired results.

The human speech production mechanism consists of the lungs, trachea (windpipe), larynx, pharyngeal cavity (throat), buckle cavity (mouth), nasal cavity, velum (soft palate), tongue, jaw, teeth and lips [7]. The lungs and trachea make up the respiratory subsystem of the mechanism. These provide the source of energy for speech when air is expelled from the lungs into the trachea. The resulting airflow passes through the larynx, which provides periodic excitation to the system to produce the voiced sounds. The three cavities of the system can collectively be termed as the main acoustic filter that shapes the sound that is generated. The velum, tongue, jaw, teeth and lips are known as the articulators. These provide the finer adjustments to generate speech. Figure 2.1 [7], shows a schematized view of the human sound production system. There are varieties of ways to produce sound. One method involves using the air pressure provided by the lungs to set the elastic vocal folds into vibratory motion. The larynx converts the steady flow of air produced by the sub glottal system into a series of puffs, resulting in a quasi-periodic sound wave. Second, periodic sounds are produced by allowing air to pass through the open glottis into the upper airway (the supralaryngeal vocal tract) where localized turbulence can be produced at constrictions in the tract. Third method involves producing transient clicks and pops by rapid release of the articulatory closure [7]. Here the sound sources arise from the local changes in the vocal tract and do not require air pressure from the sub glottal system.

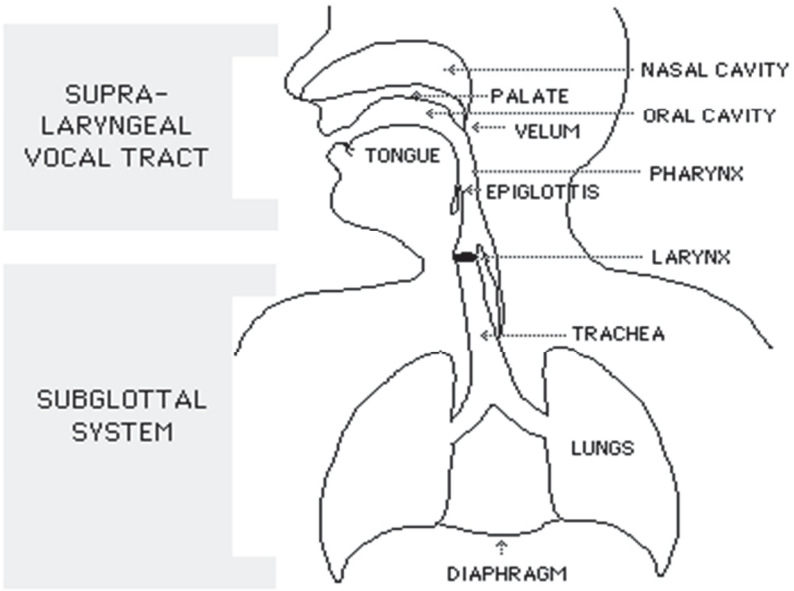

Figure 2.1: Human speech production system [7]

The excitation used to generate speech can be classified into voiced, unvoiced, mixed, plosive, whisper and silence [7]. Any combination of one or more can be blended to produce a particular type of sound. A phoneme describes the linguistic meaning conveyed by a particular speech sound. The American English language consists of about 42 phonemes, which can be classified as vowels, semivowels, diphthongs and consonants (fricatives, nasals, affricatives and whisper) as shown in Figure 2.2.

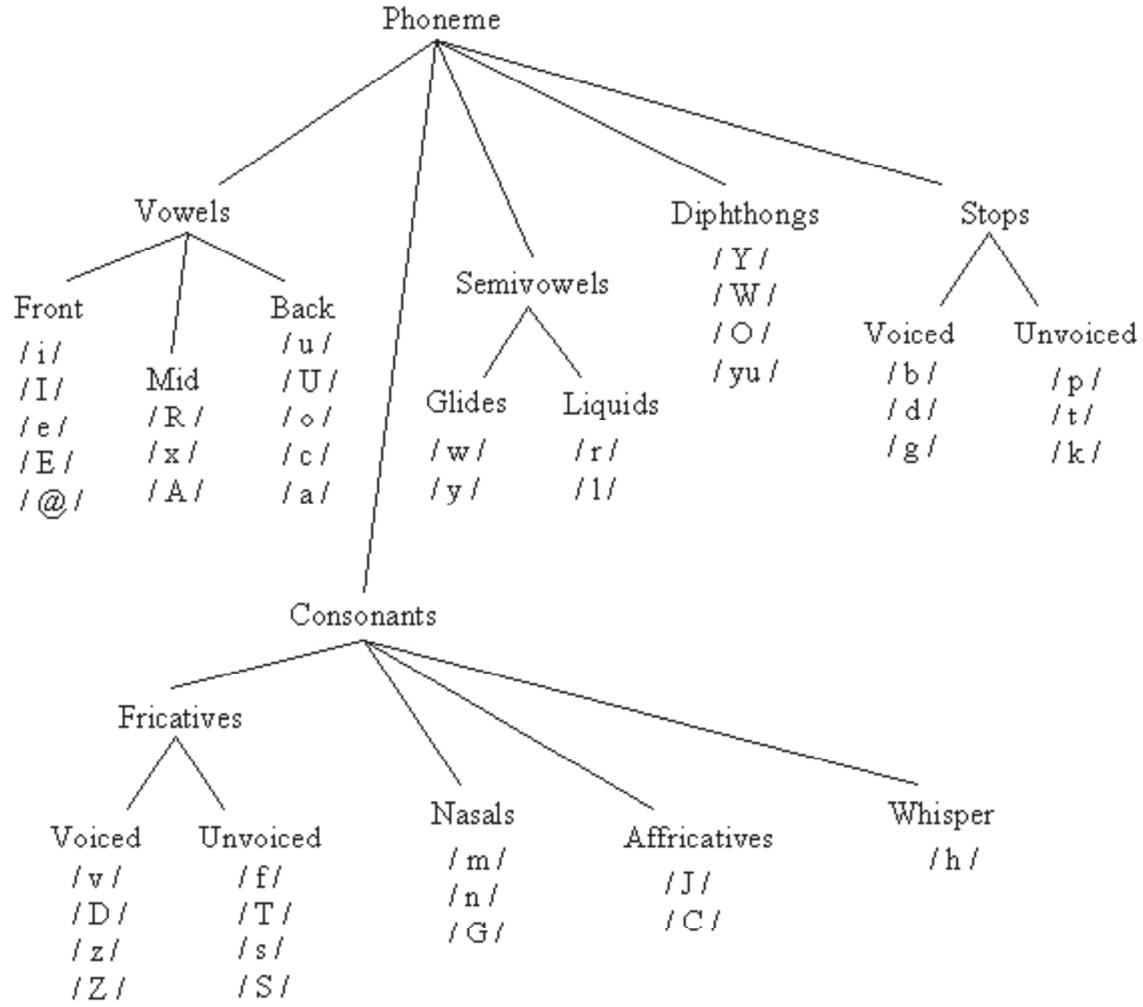

Figure 2.2 Phonemes in American English [7]

Vowels are produced due to the periodic vibrations of the vocal chords in the larynx. The frequency at which the vocal chords vibrate is called the fundamental frequency or pitch of the speech. The fricatives are caused by the turbulence of the air passing through narrow constrictions in the vocal tract, causing a random noise-like sound. Nasals are caused by acoustically coupling the nasal cavity to the pharyngeal cavity by lowering the velum. Building up pressure in front of the vocal tract and abruptly releasing it produces plosives.

The resonant frequencies generated by the vocal tract are called the formant frequencies or the formants. The formants depend on the length and shape of the vocal tract.

2.3 Filter bank:

A filter bank provides a natural decomposition of the input signal into different frequency bands [17]. These filter banks can be used to transform the time-domain input signal in to certain number of uniform frequency bands, i.e. it allows a signal to be decomposed in to several subbands, which facilitates more efficient and effective processing. This decomposition is done in frequency domain rather than in time domain, since it is having several advantages due to their computational requirements. This decomposition has vast significant benefits in various dimensions of performance they are 1) Faster convergence and lower complexity in adaptive equations [19], 2) Efficient short time spectral analysis and synthesis [7], 3) Reliable speech reorganization [8]. Due to which it finds wide variety of applications in communications, image compression, speech processing, antenna systems and digital audio industry. The transformed filter bank signals are denoted as subband signals since each of them describe a subband of original signal, which can be processed independently so that processing load can be implemented in parallel for every sub band.

The part of the filter bank that transforms a time domain input signal into a corresponding frequency domain representation is referred to as an analysis filter bank [17] and the corresponding reconstruction that transforms a frequency signal into a time signal is referred to as a synthesis filter bank [17]. An analysis-synthesis filter bank is illustrated in Figure 2.3.

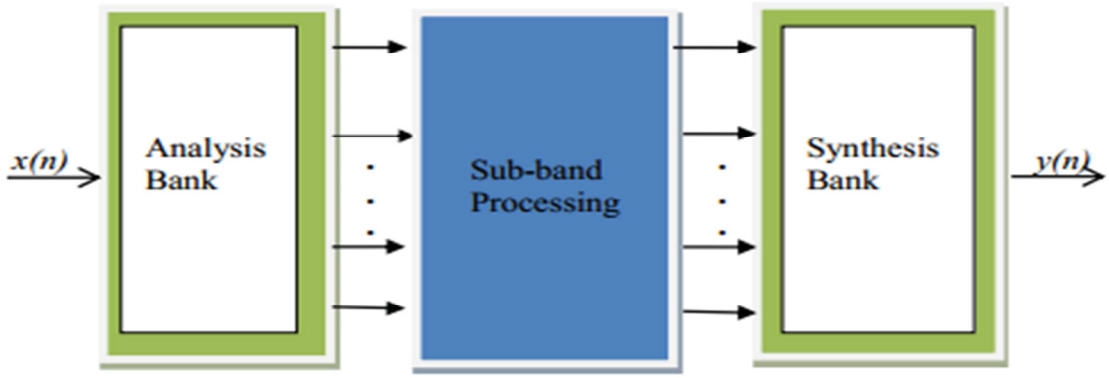

Figure 2.3 Analysis-Synthesis filter banks

2.3.1 Digital filter bank:

In general Filter Bank (FB) is defined as collection of low pass, band pass, and high pass digital filters with common input and common output. FB comprises of analysis $\{H_0(z),\ H_1(z) \cdots H_{M-1}(z)\}$ filters and synthesis filters $\{G_0(z),\ G_1(z) \cdots G_{M-1}(z)\}$ as shown

in Figure 2.4, the analysis stage splits signal into subband signals in frequency domain are called subband signals or frequency bins and synthesis stage is used to reconstruct time domain signal from frequency domain signal perfectly. As the FB performs subband decomposition, whose characteristics varies from application to application. The characteristics varies in (1) selecting FIR or IIR filters, (2) identifying time-frequency or space frequency representation, (3) designing special characteristics of analysis and synthesis stages like in defining passband deviations, transitions bands etc, and (4) also finally in designing a Perfect Reconstruction (PR) of digital signal at the output of FB. The main goal in choosing the various FB parameters is to minimize the error in reconstructed signal. In Figure 2.4, the subband processing unit introduces a transmission delay and also signals degradation due to aliasing effect. This aliasing can be reduced by using higher sampling rate than critically needed in subband and thus also reduces subband signal degradation [17]. So in this thesis we preferred in designing a uniform Discrete Fourier Transform (DFT) based filter bank with an overlapping factor, and also Fast Fourier Transform-Modulated Filter Bank (FFT-ModFB), these two approaches minimizes aliasing and magnitude/phase distortions.

The perfect reconstruction can be obtained and also aliasing can be avoided when the filters used in analysis and synthesis FB stages are ideal in case means, but in practice it is impossible to realize ideal filters. So in defining filter parameters need to choose correctly, so that aliasing can be cancelled and also output reconstruction signal $y[n]$ is almost similar to original signal $x[n]$.

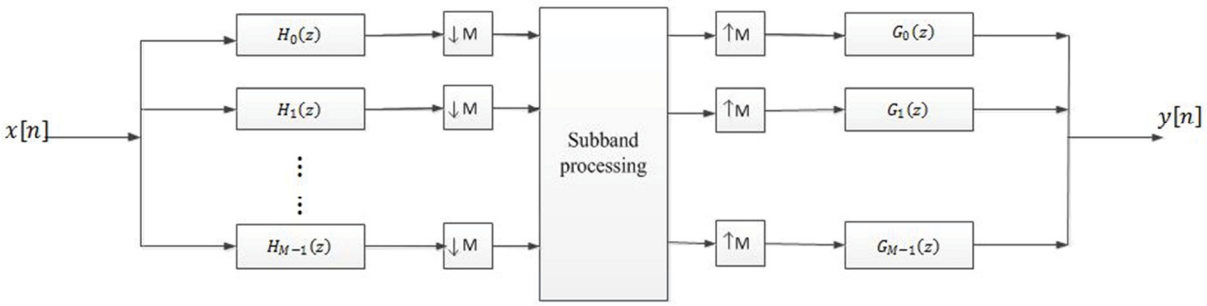

Figure 2.4 Generalized Filter Bank structure [17]

If output signal $y[n]$ is pure time delayed version of input signal $x[n]$ i.e. $y[n] = c\{x[n-n_0]\}$, $c \neq 0$ and n_0 is integer, then whose FB is having Perfect Reconstruction (PR).

2.4 Speech Enhancement

The term speech enhancement refers to methods aiming at recovering speech signal from a noisy observation. Speech enhancement techniques are concerned with algorithms that mitigate these unwanted noise effects and thus improve signal quality. The results of speech enhancement have important practical applications in systems such as mobile telephony; air-traffic radio control and voice relay networks.

Speech enhancement systems can be classified in a number of ways [9, 10] based on the criteria used or application of the enhancement system. The noise reduction systems generally can be classified based on the number of input channels (one/multiple), domain of processing (time/frequency/spatial) and type of algorithm (non-adaptive/adaptive) [10,11,12]. The speech enhancement techniques can be divided into two broad classes based on single-microphone speech enhancement and multi-microphone speech enhancement techniques [9].

The speech signal can be acquired from single or multiple channel sensors. The problem of enhancing speech degraded by additive background noise, when only a single channel is available, remains challenging due to the limitations of existing methods realistic noise conditions. One microphone input (single channel) could make speech enhancement difficult, as speech and noise are present in the same channel. Separation of the two signals would require relatively good knowledge of the speech and noise models or require that the interfering signal be present exclusively in a different frequency band than that of the speech signal. A costly solution to this problem is to use a dual microphone approach. Spatial analysis can however help immensely in speech enhancement as this gives useful information regarding the signal. In such analysis, the noise source is assumed to be statistically independent and additive. This assumption is based on the fact that most environmental noise is typically additive in nature [10].

2.4.1 Single Channel Speech Enhancement

The problem of enhancing speech degraded by additive background noise, when only a single channel is available, remains challenging due to the limitations of existing methods in realistic noise conditions. In general single channel systems constitute by depending on different statistics of speech and unwanted noise, here difficult situation is where no prior knowledge of noise is available. So the behavior of these methods depends on Signal to Noise Ratios (SNR) and the features of the noise (stationery, non-stationery, etc…).

2.4.1.1 Suppression of noise using periodicity of speech

Methods based on suppression of noise using periodicity of speech exploits the quasi-periodic nature of voiced speech. As discussed in Chapter 1, voiced speech is periodic in nature characterized by a fundamental frequency, which varies from person to person. This technique however, depends heavily on the accurate estimation of the pitch period (inverse of the pitch) of the speaker's voice. A simple method based on this criterion is the adaptive comb filter [9]. In this method, a series of notch filters are used so as to filter out any spectral content between the fundamental frequency and its harmonics. Another method is the single channel adaptive noise cancellation technique [13]. In this method, a delayed version of the speech signal is used as the input to an adaptive Least Mean Square (LMS) [7] filter while the input used as the reference signal. The delay decorrelates the noise in the input signal with that present in the reference. When the delay is equal to an estimate of the pitch period, there is a correlation in the speech content of the two signals. A major disadvantage of these methods is that there is no improvement in quality of the unvoiced speech portions. Moreover, an accurate pitch extraction algorithm is crucial to achieving good performance.

2.4.1.2 Model-based speech enhancement

Methods based on model-based speech enhancement are also called statistical – model based methods [24]. These methods are usually used when there is no knowledge of the statistical properties of the speech or noise signal. Speech production models like Auto Regressive – Moving Average (ARMA), Auto Regressive (AR) or Moving Average (MA) are used instead [7, 24]. This involves the estimation of the speech model parameters and then the estimation of the enhanced signal by re-synthesis using speech model parameters or by using a Wiener filter or Kalman filter [12]. The Wiener filter is a popular adaptive technique that has been used in many enhancement methods. The basic principle of the Wiener filter is to estimate an optimal filter from the noisy input speech by minimizing the Mean Square Error (MSE) between the desired signal $s(k)$ and the estimated signal $\hat{s}(\mathrm{k})$ i.e. $E[(s(k) - \hat{s}(\mathrm{k}))^2]$. The Wiener filter can be given in the frequency domain by:

$$H(w) = \frac{P_s(\omega)}{P_s(\omega)+P_n(\omega)} \qquad (2.1)$$

Where $P_s(\omega)$ is the Power Spectral Density (PSD) of the speech and $P_n(\omega)$ is the PSD of the noise spectrum calculated during periods of non-speech activity. From equation (2.1) it is obvious that a priori knowledge of the speech and noise power spectra is necessary. The speech power spectrum is estimated using the estimated speech model parameters [6].

2.4.1.3 Short-term spectral amplitude techniques

These algorithms are primarily based on Short Time Spectral Attenuation (STSA). Widely researched and applied examples of STSA speech enhancement are spectral subtraction as originally proposed by Berouti et.al. [22] and the Ephraim-Malah MMSE short-time spectral amplitude estimator [22]. These methods can be viewed in terms of applying a spectral gain to each frequency bin in a short-time frame of the noisy speech signal. Since the spectral components are assumed to statistically independent, the gain is adjusted individually as a function of the relative local SNR at each frequency. With low SNR regions attenuated relative to high SNR regions, an improvement in the overall SNR is achieved. A good estimate of the instantaneous noise spectrum is crucial in the estimation of the local SNR, without which quality would degrade due to the presence of either high residual noise or high speech distortion in the enhanced speech.

A general representation of the technique is given in Figure 2.5. The input to the system is the noise-corrupted signal *x(n)*. While there are many methods for the analysis-synthesis processing, the Short Term Fourier Transform (STFT) of the signal with Overlap and Add (OLA) [4] is the most commonly used method. The spectral amplitude $|X(k)|$ of the noisy input signal *x(n)* is modified using a correction factor. Usually this correction factor could be the spectral amplitude of the estimated noise signal *d(n)* , measured during periods of silence/non-activity in the speech signal or obtained from a reference channel (dual-channel method). The correction is obtained by subtracting the spectral amplitude of the noise signal from that of the noisy speech input. Hence, these methods are also referred to as subtractive-type algorithms. If the noise is assumed to be uncorrelated with the speech signal, then the corrected amplitude can be considered as an estimate $|\hat{s}(\mathrm{k})|$ of the original clean speech signal $s(n)$. The unprocessed phase of the noisy input signal is used to synthesize the enhanced speech signal under the assumption that the human ear is not able to perceive the distortions in the phase of the speech signal.

Spectral subtraction is a well-known noise reduction method based on the STSA estimation technique. The basic power spectral subtraction technique, as proposed by Boll [3], is popular due to its simple underlying concept and its effectiveness in enhancing speech degraded by additive noise. The basic principle of the spectral subtraction method is to subtract the magnitude spectrum of noise from that of the noisy speech. The noise is assumed to be uncorrelated and additive to the speech signal. An estimate of the noise signal is measured during silence or non-speech activity in the signal. Since Boll [3] first proposed this

method, several variations and enhancements have been made to this technique to overcome some inherent drawbacks such as musical noise in the method.

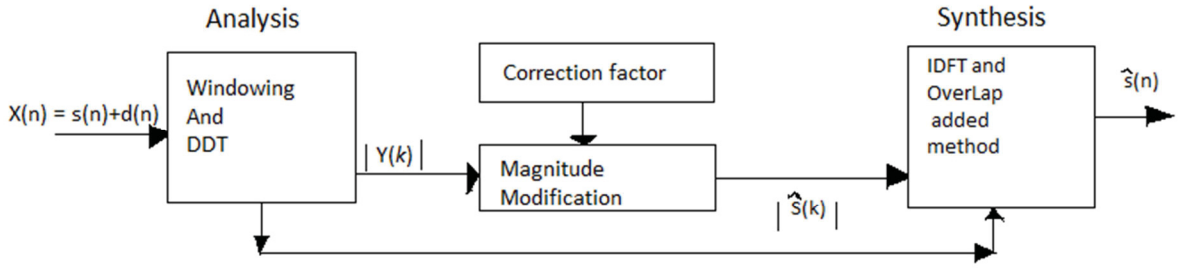

Figure 2.5 Short-time spectral magnitude enhancement system

2.5 Spectral Subtraction

The Spectral Subtraction algorithm is considered to be one of the first algorithms proposed for noise suppression [3]. Many variations for spectral subtraction algorithms are developed aiming at better suppression of noise. It is based on simple principle, assuming the background noise is acoustically added to clean speech; the estimate of clean speech signal can be obtained by subtracting the noise magnitude spectrum from noisy speech spectrum. Spectral information required to describe the noise spectrum is obtained during the periods when the speech signal is absent. The IDFT of the estimated signal spectrum using the phase of the noisy signal gives the enhanced output signal. The algorithm is computationally simple as it involves only Fourier transform both inverse and forward approach for suppression of acoustic noise in order to improve the intelligibility and quality of processed speech signal. Section 2.5.1, presents the basic principle of the technique, Section 2.5.2 discusses the drawbacks in the method.

2.5.1. Principle of Spectral Subtraction

Let us assume that $x(n)$ be noise corrupted signal, which is composed of clean speech signal $s(n)$ and undesired additive noise signal $d(n)$ i.e.

$$x(n)=s(n) + d(n) \qquad (2.2)$$

Here noise is considered to be stationary or slowly varying process. Taking the DTFT on both sides of equation 2.2 we get

$$X(\omega) = S(\omega) + D(\omega) \qquad (2.3)$$

The polar form of equation (2.3) is given as

$$X(\omega) = |X(\omega)|e^{j\theta_x(\omega)} \quad (2.4)$$

where $|X(\omega)|$ is the magnitude spectrum and $\theta_x(\omega)$ is the phase of noisy signal. Similarly for noise spectrum the polar form is defined as below

$$D(\omega) = |D(\omega)|e^{j\theta_d(\omega)} \quad (2.5)$$

$|D(\omega)|$ is magnitude of noise spectrum generally unknown but it can be replaced with average value computed during non-speech activity. Since we know that phase does not affect speech intelligibility [15] and have little effect on speech quality, the phase of noisy spectrum $\theta_x(\omega)$. Applying these assumptions to equation (2.3) we get an estimate of clean signal spectrum

$$\hat{S}(\omega) = (|X(\omega)| - |\hat{D}(\omega)|)e^{j\theta_x(\omega)} \quad (2.6)$$

where $|\hat{D}(\omega)|$ is the estimated average magnitude spectrum of noise measured during non-speech activity. The enhanced speech signal can be obtained by taking IDFT of $\hat{S}(\omega)$. The general form of the spectral subtraction algorithm is shown in Figure 2.6 [21].

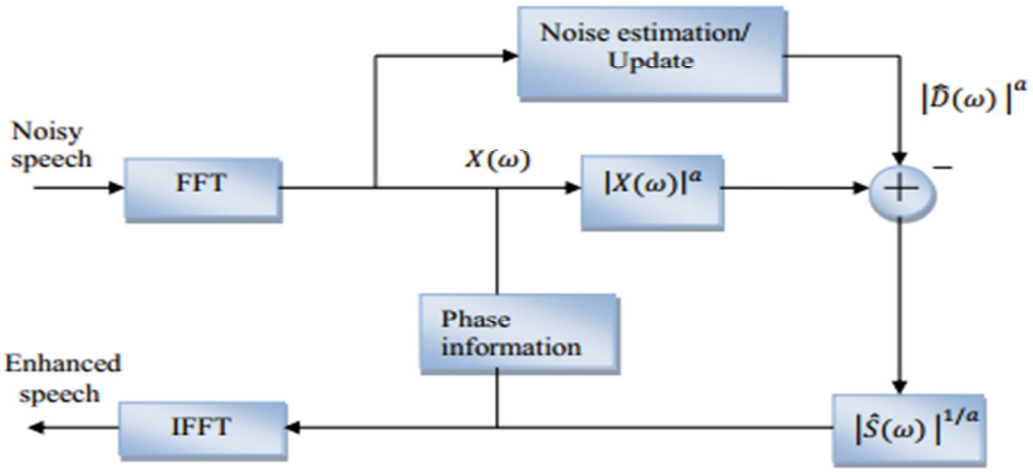

Figure 2.6 General form of spectral Subtraction [21]

In the Figure 2.6, a block contains Noise estimation/Update whose common method of noise estimation involves the use of Voice Activity Detector (VAD) to detect pauses in speech. The noise estimate is then obtained by recursively smoothened adaptation of noise during the detected pauses. In stationery background noise, such an estimator is generally reliable. However non-stationery noise, with noise spectrum levels changing in time, cannot be tracked adequately by a recursive noise estimation method that adapts only during detected speech pauses. This is especially true of environmental noises such as factory or battlefield noise that are characterized by large, irregular random bursts embedded in a relatively stationery background, Even if VAD is reliable changes in the noise spectrum occurring during active speech cannot influence the noise estimate in a timely manner. Due to

difficulty in tracking highly non-stationery noise spectra, VAD based algorithms are effective only in suppressing the stationery noise component generally leaving noise bursts unattenuated in enhanced speech. Now we will be limited to single channel enhancement techniques, as these are the most common types of enhancement systems found in many applications.

2.5.2 Drawbacks of Spectral Subtraction

The subtraction process leads to the speech distortion if it is not done carefully. The speech terms might be removed if too much is subtracted and if two little is subtracted then much of noise remains which affect the quality of speech. The equation (2.6) can be negative owing to inaccuracies in the noise estimations. This can be eliminated by half wave rectification process i.e. set the negative spectral components to zeros.

$$|\widehat{S}(\omega)| = |X(\omega)| - |\widehat{D}(\omega)| \quad if \quad |X(\omega)| > |\widehat{D}(\omega)|$$
$$= 0 \qquad otherwise \qquad (2.7)$$

Many methods have been proposed to reduce the speech distortions over subtracting estimates of noise spectrum [22, 24]. Setting of negative values to zero in equation (2.7) is a nonlinear processing which creates small peaks in the magnitude spectrum occurring randomly in time and frequency. In time domain it introduces new type of noise called musical noise [22], which sounds like tones with frequencies that change randomly from frame to frame. While the spectral subtraction method is easily implemented and effectively reduces the noise present in the corrupted signal, there exist some glaring shortcomings, which are given below:

Residual noise (musical noise)

It is obvious that the effectiveness of the noise removal process is dependent on obtaining an accurate spectral estimate of the noise signal. The better the noise estimate, the lesser the residual noise content in the modified spectrum. However, since the noise spectrum cannot be directly obtained, we are forced to use an average estimate of the noise. Hence there are some significant variations between the estimated noise spectrum and the actual noise content present in the instantaneous speech spectrum. The subtraction of these quantities results in the presence of isolated residual noise levels of large variance. This residual spectral content manifest themselves in the reconstructed time signal as varying tonal sounds resulting in a musical disturbance of an unnatural quality. This musical noise can be even more disturbing and annoying to the listener than the distortions due to the original noise content. Several residual noise reduction algorithms have been proposed to combat this

problem. However, due to the limitations of the single-channel enhancement methods, it is not possible to remove this noise completely, without compromising the quality of the enhanced speech. Hence there is a trade-off between the amount of noise reduction and speech distortion due to the underlying processing.

Roughening of speech due to the noisy phase

The phase of the noise-corrupted signal is not enhanced before being combined with the modified spectrum to regenerate the enhanced time signal. This is due to the fact that the presence of noise in the phase information does not contribute immensely to the degradation of the speech quality. This is especially true at high SNRs (> 5 dB). However at lower SNRs (< 0 dB), the noisy phase can lead to a perceivable roughness in the speech signal contributing to the reduction in speech quality. Experiments conducted by Schroeder [14] have corroborated this fact. Estimating the phase of the clean speech is a difficult and will greatly increase the complexity of the method. Moreover, the distortion due to noisy phase information is not very significant compared to that of the magnitude spectrum, especially for high SNRs. Hence the use of the noisy phase information is considered to be an acceptable practice in the reconstruction of the enhanced speech signal.

Most speech enhancement algorithms, including the spectral subtraction methods, try to optimize noise removal based on mathematical models of the speech and noise signals. However, speech is a subtle form of communication and is heavily dependent on the relationship of one frequency with another. Hence, while conventional speech enhancement algorithms can increase the speech quality of the noisy speech by increasing the SNR, there is no significant increase in speech intelligibility. Algorithms should take into account the subtleties of speech and incorporate methods based on the perceptual properties of the speech signal. The spectral subtraction methods, as well as most other methods, suffer from this drawback. Studies [4, 10] have shown that there is no improvement in the intelligibility in the speech signals enhanced by the spectral subtraction method.

Several variants of the spectral subtraction method originally proposed by Boll [15] have been developed to address the problems of the basic technique, especially the presence of musical noise. A variety of pre-processing and post-processing methods have also proved to help reduce the presence of musical noise while minimizing speech distortion. In this thesis the spectral subtraction is implemented based on Minimum Statistics which is described clearly in chapter 4 where we extent to maximum we reduce musical noise.

CHAPTER-3

3. DESIGN, IMPLEMENTATION AND EVALUATION OF FILTER BANKS

In designing of filter bank, Fourier analysis and synthesis plays a vital role for analyzing and modeling of quasi-stationary signals, such as speech signals [16]. So in this thesis, the implementation of filter bank is based on the Short Time Fourier Transforms (STFT), regarding this early theoretical framework for analysis and reconstruction using STFT was formulated in [17]. Since then time-frequency systems based on decimated filter banks can be viewed has the generalization of the STFT representation.

3.1 Design of WOLA filter bank:

The Weighted Overlap Filter Bank [WOLA-FB] is highly efficient implementation of an over-sampled DFT bank, offers a low computationally cost with effective lower delay, perfect/or near reconstruction system [18, 19]. The choice in choosing the WOLA-FB parameters has an effect on aliasing, frequency and time resolution, and finally group delay. WOLA-FB has two stages they are analysis and synthesis stages. The particular care has been taken in selecting filter parameters such as analysis window function $w(n)$, length of analysis window *N*, number of subbands *K*, decimation rate D_F.

3.1.1 An analysis bank based on DFT:

WOLA structure is block based transform interpretation, whose simplified block diagram of analysis bank is shown in Figure 3.1 [20]. In the analysis stage the input signal is shifted in *D* samples at a time into input buffer $u[n]$ of length *N* samples. This input buffer $u(n)$ is windowed with a prototype FIR filter window function of hamming or hanning, of length *N* and stored into temporary buffer $t_1(n)$ of length *N* samples, i.e. $t_1(n) = u(n)w(n)$. This buffer $t_1(n)$ is then time shifted into a vector $t_2(n)$ and then circularly shifted by *K/2* samples, to produce a zero-phase signal for DFT. The signal is transformed into frequency domain by DFT so output of analysis FB provide both magnitude and phase information, i.e. they are in complex form. By doing *K* size modulo FFT operation we generate *K* number of subband signals are $X_0(\lambda)$, $X_1(\lambda), \cdots X_{K-1}(\lambda)$, this are the output signals of analysis stage. Here in the Figure 3.1, the terms $g_0, g_1, \cdots g_{K-1}$ are subband gain processing function, this gain functions of speech enhancement processing algorithms.

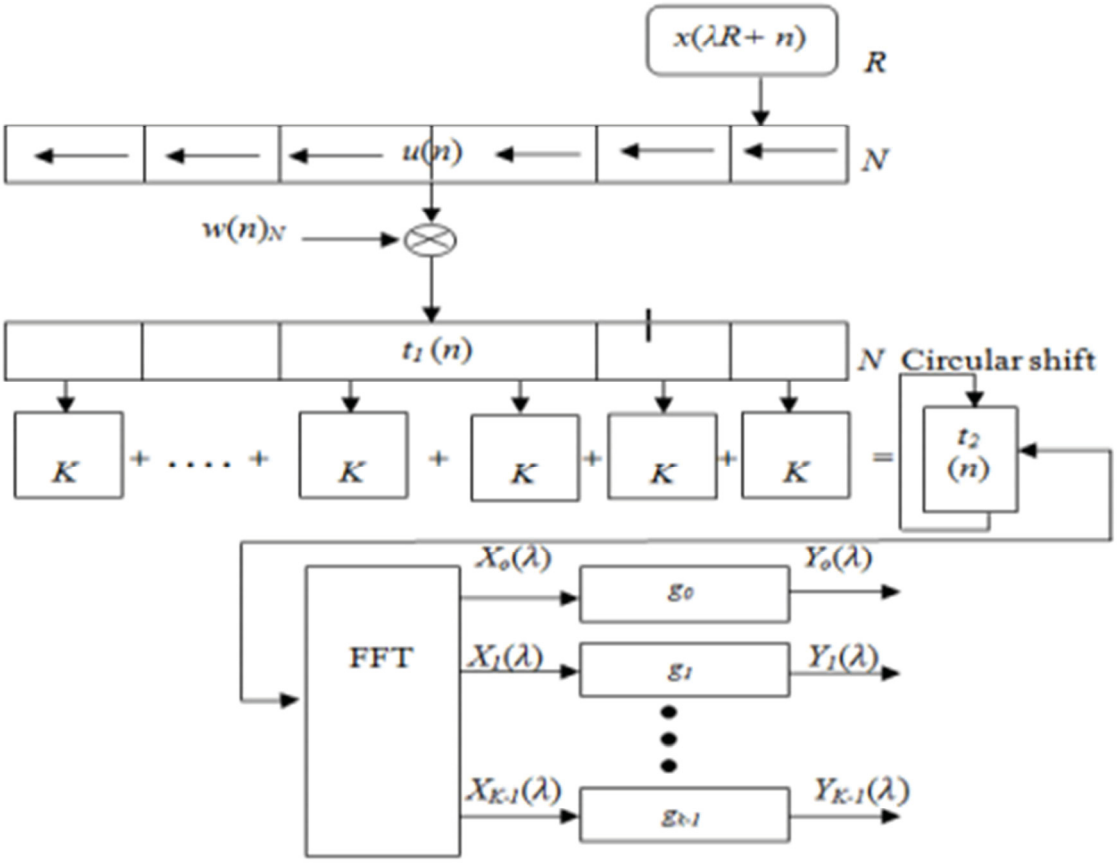

Figure 3.1 Analysis stage of WOLA filter [20]

3.1.2 A synthesis bank based on inverse DFT:

The generalized synthesis structure of WOLA filter bank is shown in Figure 3.2 [20]. Here in this stage only actual WOLA procedure is undergone. The subband processed signal $Y_0(\lambda)$, $Y_1(\lambda), \cdots Y_{K-1}(\lambda)$ here K stands number of subbands and λ is bin index $\lambda \in 0,1, \ldots, N-1$, these are obtained after applying the noise suppression algorithm, then transformed to time domain signal by applying K size IFFT. To counteract the circular operation in analysis bank, the transformed signals undergo circular shifting operation by $K/2$ samples and stored in temporary buffer $t_2(n)$. Then this buffer $t_2(n)$ is stacked by repetition into buffer $t_4(n)$ of length N/R, where N is window length and R is decimation factor. This buffer $t_4(n)$ is weighted by synthesis window function $z(n)$ of size N/R, then weighted data is summed with the data in buffer $t_5(n)$ of length N/R, and finally output buffer $t_5(n)$ data is over written with summation result i.e. $t_5[n] = t_5(n) + z(n).t_4(n)$, this is procedue is known as WOLA procedure. Then the resulted data is shifted by R samples resulting into output buffer $y(\lambda R + n)$ and adding zeros of R samples in buffer $t_5(n)$ and then the procedure is repeated for next block of input samples as shown in Figure 3.2

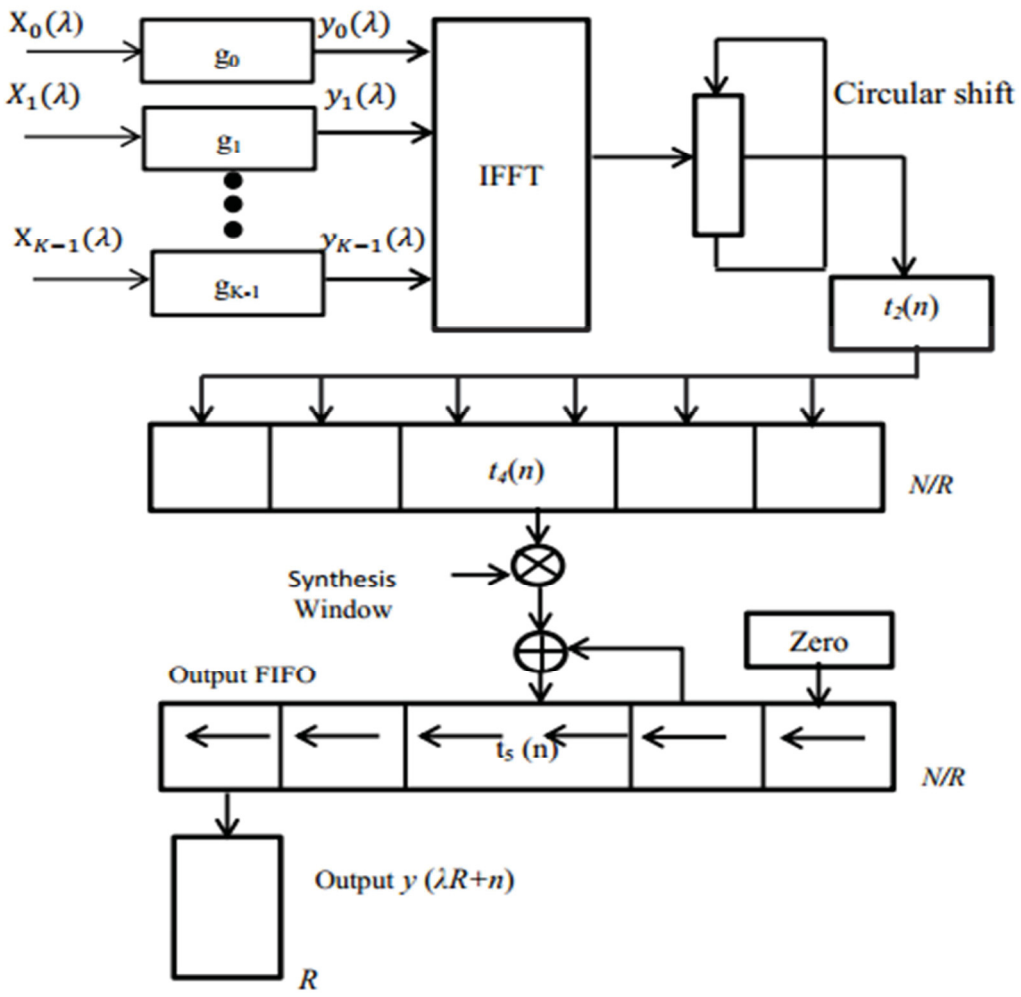

Figure 3.2 Synthesis stage of WOLA filter [20]

3.1.3. Implementation and Evaluation of Filter Bank

As said earlier WOLA-FB is a more efficient filter bank technique when defining by four variables namely *L* the length of analysis window, *R* the decimation rate, *K* the number of subbands, along with the analysis window function *w(n).* So now in our implementation we have selected the length of analysis window as *L*=256, number of subbands *K*=128, block rate *R*=64, synthesis window decimation rate D_F=1

The WOLA filter bank is implemented and it is tested with speech signal sampled at sampling rate of 8 kHz, the evaluation outputs are presented in this section. The Figure 3.3 shows the reconstructed pure speech over the original, when original signal is passed through filter bank alone, the reconstructed signal whose PESQ quality score is 4.301 out of 4.500 MOS, so this WOLA-FB is lossless filter bank in reconstruction of signal.

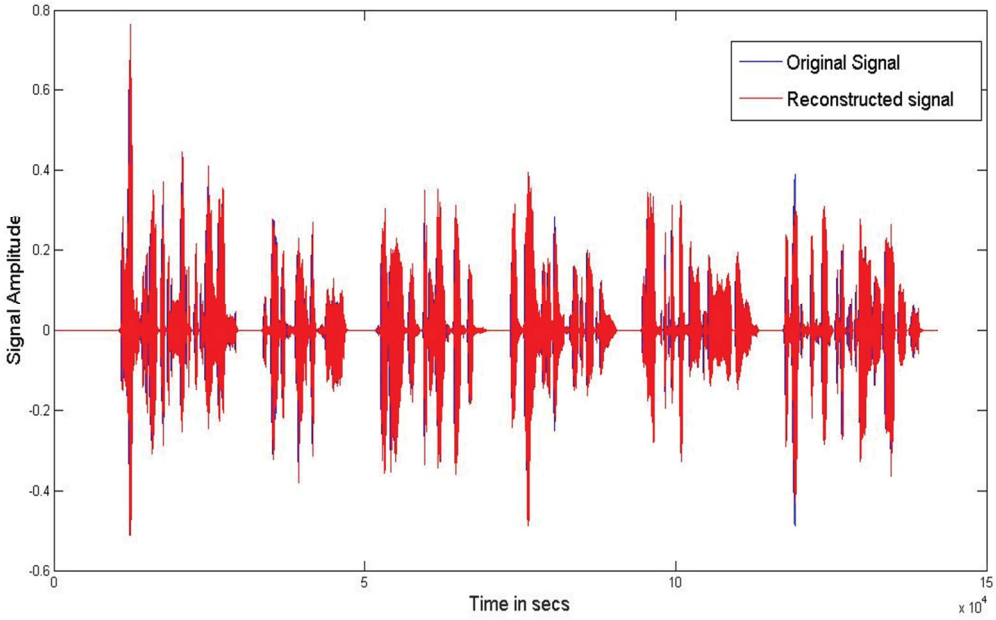

Figure 3.3 Original and Reconstructed Speech signal

We can say that the speech signal is perfectly reconstructed by the WOLA filter bank without any gain functionality of enhancement algorithm.

3.2. Design of FFT-Modulated Filter Bank

Among general multirate filter banks, DFT modulated filter banks have gained much popularity due to their ease in design and realization. An M-channel DFT modulated filter bank can be characterized by one or two Prototype Filters (PF) and efficiently implemented by polyphase structure. Therefore, the design of DFT modulated filter bank only involves PFs and the design complexity is considerably reduced compared with that of the general filter bank.

We use a set of *K* bandpass filters $H_k[z]$ with impulse response functions $h_k[n]$ each of length *N* taps. These bandpass filters are created by modulating (frequency shifting) a low-pass prototype filter (which is equal to the first band pass filter at DC frequency, i.e. $h_o[n]$ according to

$$h_k[n] = W_k^{-kn} h_0[n], \text{ for } n = 0 \dots \dots N-1, \quad (3.1)$$

where $W_k = e^{-j\frac{2\pi}{k}}$. The Z-transform of each modulated bandpass filter is

$$H_k[z] = \sum_{n=0}^{\infty} h_k[n] z^{-n} = \sum_{n=0}^{\infty} W_k^{-kn} h_0[n] z^{-n} = \sum_{n=0}^{\infty} h_0[n] \, (W_K^k z)^{-n} = H_0[W_K^k z]. \quad (3.2)$$

So, it is clear that each bandpass filter $H_k[z] = H_0[W_K^k z]$ is a frequency shifted (modulated) version of the prototype filter $H_o[z]$. The notation "uniform" in the section title

comes from the fact that the filters are uniformly distributed on the frequency axis during the modulation process. There are also non-uniform modulation techniques which are not described here.

We shall now assume that the band pass signals, i.e. the input signal filtered by each modulated band pass filter, are subject to decimation a factor D, where $D=K/O$, and O denotes the over-sampling ratio. A filter with a following decimator can be implemented using a polyphase composition [18]. The corresponding polyphase implementation is achieved by dividing the prototype filter $H_o[z]$ into D polyphase components, denoted as $P_d[z]$ with impulse response function $p_d[n]$ for d=0, 1,...D−1, according to

$$H_0[z] = \sum_{d=0}^{D-1} z^{-d}\, P_d[z^D], \qquad (3.3)$$

$$p_d[n] = \; h_0[d+nD], \qquad (3.4)$$

$$P_d[z] = \sum_{n=0}^{\infty} p_d[n]\, z^{-n} = \sum_{n=0}^{\infty} h_0[d+nD] z^{-n}. \qquad (3.5)$$

Hence inserting equation (3.3) and (3.5) into equation (3.2), yields a polyphase representation of each modulated bandpass filter $H_k[z]$ as follows

$$H_k[z] \;=\; H_0\left[W_K^k z\right] \;= \sum_{d=0}^{D-1} W_K^{-dk} z^{-d} P_d[\, W_K^{kD} z^D]$$

$$= \sum_{d=0}^{D-1} W_K^{-dk} z^{-d} \sum_{n=0}^{\infty} h_0[d+nD](W_K^{kD} z^D)^{-n} \qquad (3.6)$$

$$= \sum_{d=0}^{D-1} W_K^{-dk} z^{-d} \sum_{n=0}^{\infty} h_0[d+nD] W_K^{-nkD} z^{-nD} = \sum_{d=0}^{D-1} W_K^{-dk} z^{-d} \sum_{n=0}^{\infty} h_0[d+nD] W_O^{-nk} z^{-nD},$$

where we used that $W_K^{-nkD} = W_O^{-nk}$. The factor W_O^{-nk} is used to introduce O groups of subband polyphase components, where each group will be described by the index g=0, 1...O−1. Each group g contains subbands having indices k=g+lO for l =0, 1...D−1. Considering only the indices k belonging to group g, i.e. k=g+lO, yields $W_O^{-nk} = W_O^{-n(g+lO)} = (W_O^g)^n$. Each subband component of group g is expressed as

$$H_g + lO[z] \;= \sum_{d=0}^{D-1} W_K^{-d(g+lO)} z^{-d} . \sum_{n=0}^{\infty} h_0[d+nD] W_O^{-n(g+lO)} z^{-nD}$$

$$= \sum_{d=0}^{D-1} W_D^{-dl} z^{-d} . \sum_{n=0}^{\infty} W_K^{-dg}\, h_0[d+nD](W_O^g)^n z^{-nD} \;=$$

$$= \; \sum_{d=0}^{D-1} W_D^{-dl} z^{-d} Q_{g,d}[z^D] = \left(W_D^{-0l} \ldots W_D^{-(D-1)l}\right) . \begin{pmatrix} z^{-0} Q_{g,0}[z^D] \\ \vdots \\ z^{-(D-1)} Q_{g,D-1}[z^D] \end{pmatrix}, \qquad (3.7)$$

where we used the fact that $W_D^{-dlO} = W_D^{-dl}$. We denote the filters $q_{g,d}[n] = W_k^{-dg}(W_O{}^g)^n h_o[d+nD]$ as the d^{th} modified polyphase component of group g, and they have the Z-transforms $Q_{d,g,}[z] = \sum_{n=0}^{\infty} q_{g,d}[n]z^{-1}$. We now stack all subbands related to the g^{th} group, i.e. subbands with indices $k=g+lO$ for l =0, 1...D−1, in a vector $H_g[z]$ according to

$$H_g[z] = \begin{pmatrix} H_g[Z] \\ H_{g+O}[z] \\ \vdots \\ H_{g+(D-1)O}[z] \end{pmatrix} = \begin{pmatrix} W_D^{-00} & \cdots & W_D^{-(D-1)0} \\ \vdots & \ddots & \vdots \\ W_D^{-0(D-l)} & \cdots & W_D^{-(D-1)(D-l)} \end{pmatrix} \cdot \begin{pmatrix} z^{-0}Q_{g,0}[z^D] \\ \vdots \\ z^{-(D-1)}Q_{g,D-1}[z^D] \end{pmatrix} = DW_D^{-1} \begin{pmatrix} z^{-0}Q_{g,0}[z^D] \\ \vdots \\ z^{-(D-1)}Q_{g,D-1}[z^D] \end{pmatrix}, \tag{3.8}$$

where W_D is identified as the D^{th}-order Discrete Fourier Transformation (DFT) matrix, i.e. W_D^{-1} is the D^{th}-order Inverse Discrete Fourier Transformation (IDFT) matrix. Note that the subband indices in each group are reordered and thus need to be rearranged in an increasing order after each analysis step. Note that the size of the DFT/IDFT operations is D, the decimation rate. This filter bank structure requires that a block of D new data samples are input to the filter bank for each update of the filter bank states. Hence, the sampling rate of the subband signals is f_s/D, where f_s is the sampling frequency of the full band signal $x[n]$. If D is a power of two, e.g., $D=2p$ for any integer $p>0$, then the DFT/IDFT operation is efficiently implemented using the Fast Fourier transform (FFT) and its inverse, the IFFT. Many DSP's have built in support for FFT/IFFT which may be exploited to yield an efficient implementation. The analysis stage of this filter bank is illustrated in Figure 3.4.

The synthesis of this filter bank is illustrated in Figure 3.5. It may, however, be noted that the major difference between analysis and synthesis lies in the filters $q_{g,d}[n]$, of the analysis, that are replaced by $r_{g,d}[n]$ in the synthesis. These filters $r_{g,d}[n]$ are referred to as the d^{th} modified polyphase component of group g, but they are based on a different prototype filter. It is also noted that the subband signals of the FFT modulate filter bank are complex valued data, which may require special attention in some implementations

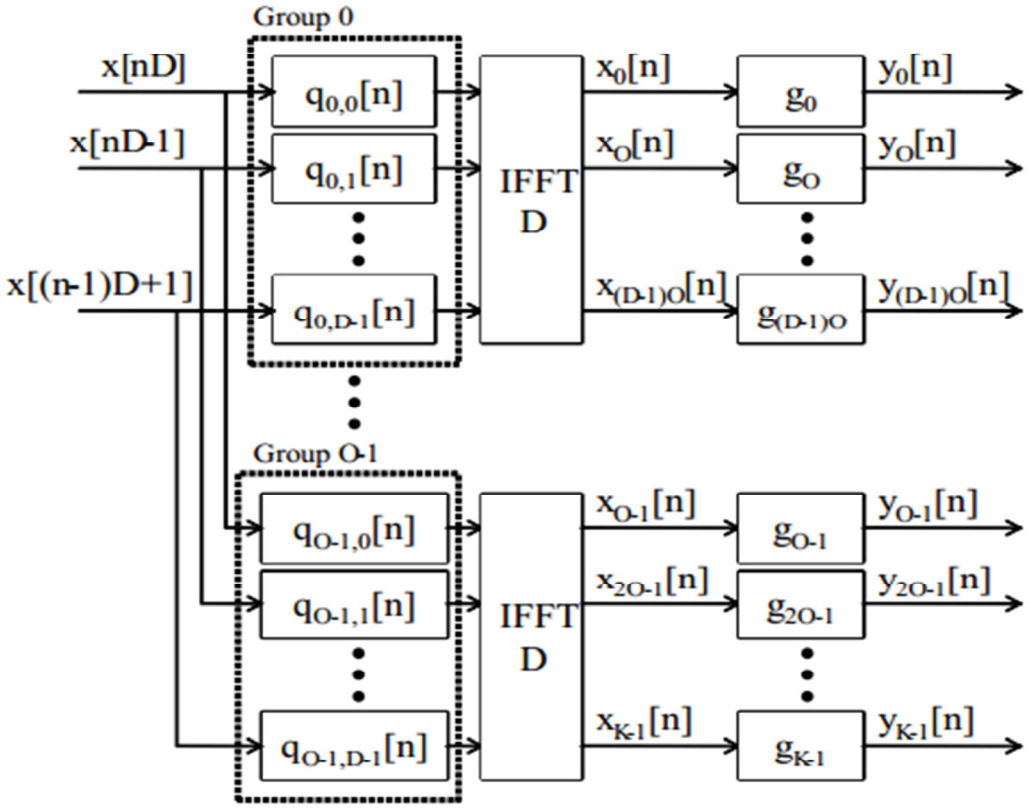

Figure 3.4 Uniform FFT Modulated Analysis Filter Bank [20]

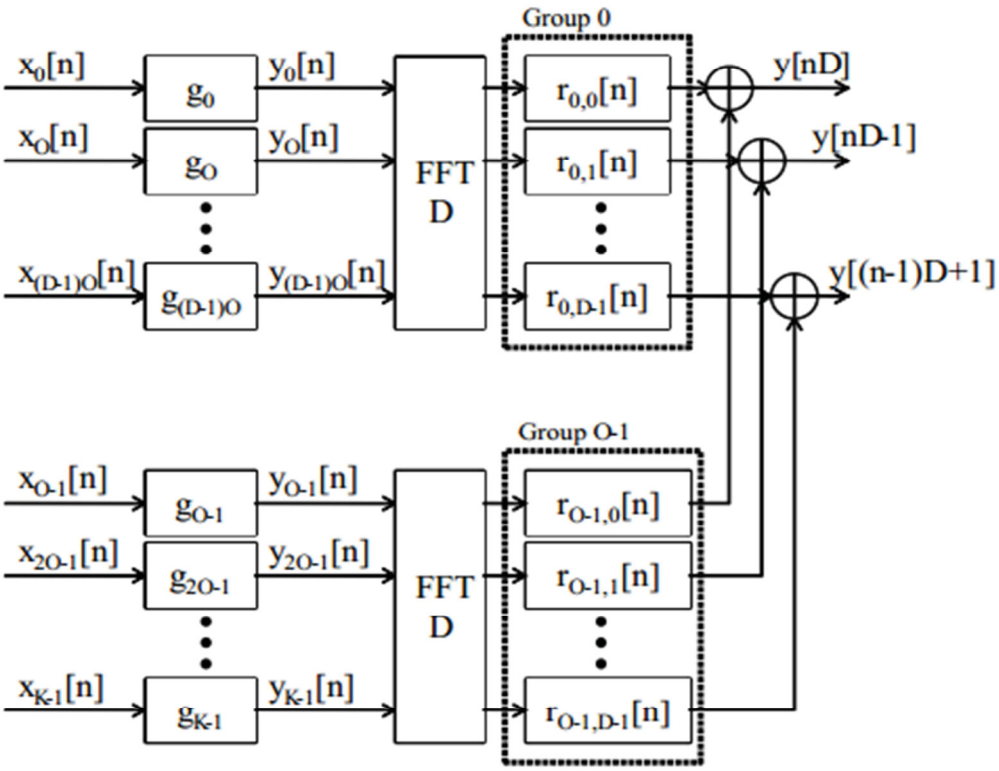

Figure 3.5 Uniform FFT Modulated synthesis FB stage [20]

3.2.2 Implementation and Evaluation of Filter Bank

In implementation of FFT Mod-FB the following parameters such as length of analysis window (L) as, number of subbands (K), block rate (R), synthesis window decimation rate (D_F), number of PF components (P) are taken into consideration carefully so that FB can reconstruct the signal without any loss due to transformation from time to frequency and vice-versa. So now in our implementation we have selected as L=256, K=128, R=64, D_F=1 and P=4.

The FFT Modulated FB is implemented and evaluated with 8 kHz pure speech signal, the output of filter bank in reconstruction of signal without any processing of gain function of noise suppression algorithm is presented in this section. The Figure 3.6 shows the reconstructed pure speech over the original, when original signal is passed through filter bank alone, the reconstructed signal whose PESQ quality score is 4.371 out of 4.500 MOS, so this FFT modulated is lossless FB in reconstruction of signal.

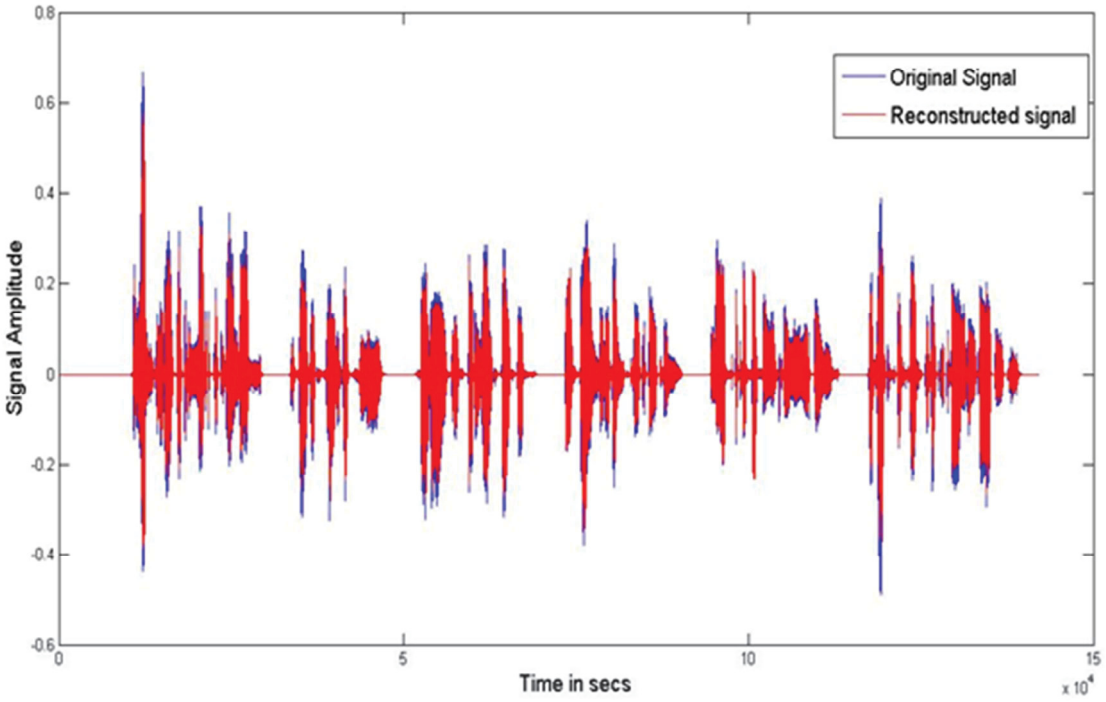

Figure 3.6 Original and Reconstructed Speech signal using FFT Modulated FB

CHAPTER-4

4. DESIGN AND IMPLEMENTATION OF MINSSS

In general the spectral subtraction method involves noise estimation using the VAD algorithm. The noise estimation is the most important factor in enhancing the speech signal. If the estimate of the noise is too low, residual noise is introduced and if the estimate of noises is too high we face a problem with a decrease in intelligibility due to distorted speech components in the signal. The VAD algorithm is used to estimates the noise spectrum during the absence of speech. The whole process runs on frame-by-frame basis, where each frame may last 20-40 msec. If a segment of speech contains voice active (VAD=1) and if there is no speech it is assumed as noise (VAD=0). This process works only to remove stationary noise but in real environment we come across various kind of noise (babble) whose spectral characteristics are not constant [25]. To overcome this issue Martin [21] has come with a unique solution known as minimum statistics. The Spectral Subtraction based on Minimum Statistics is one of the influential method for speech enhancement, which is usually able to track non stationary noise signals [21]. The problem of conventional spectral subtraction method is the requirement of speech activity detector during noise power estimation [15] which increases computational complexity while spectral subtraction based on minimum statistics uses a finite window of sub-band noise power to estimate the noise power.

4.1. Spectral Subtraction using minimum statistics

We have selected this algorithm because this needs no additional equipment due to its simplicity. The block diagram of the spectral subtraction method based on minimum statistics is shown in Figure 4.1 [21]. This algorithm uses DFT filter bank for the analysis of disturbed speech signal, and modifies the short time spectral magnitude to make the synthesized signal as close to desired speech signal. The SNR of each sub-band is calculated by using estimated noise power to control the over subtraction factor and this factor reduces the residual noise. The subtraction rule is designed by using estimated noise power with over subtraction factor for computing the optimal weighting of spectral magnitudes. It is a technique which addresses the problem of noise power estimation by essentially eliminating the need for voice activity detectors without a substantial increase in computational complexity. The algorithm is capable of tracking non-stationary noise during the speech activity. The algorithm divides the signal into small segments and transforms into short-time subband signal power. The short-time subband signal power estimate of the noisy

signal exhibits distinct peaks and valleys, where peaks determines the speech activity and the valley of the smoothed noise estimate gives the estimate of subband noise power. In addition, algorithm eliminates residual noise by taking the over subtraction as a function of subband SNR. Based on the obtained over subtraction factor and the noise power estimate we get the optimal weighting of spectral magnitude.

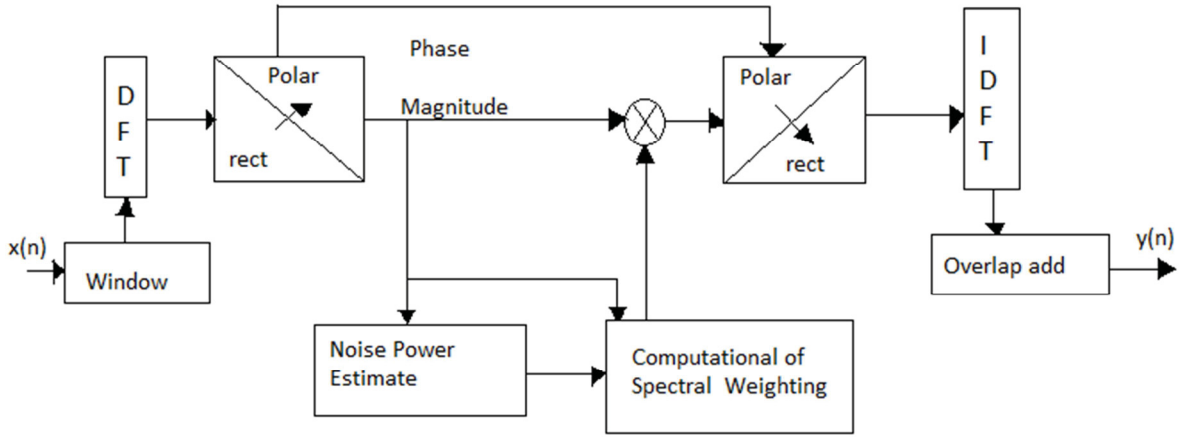

Figure 4.1: Block diagram of Spectral Subtraction by using Minimum Statistics.

4.1.1. Description

Let us take a speech $s(n)$ and noise $d(n)$ signal with a mean as zero, which are additive in nature resulting into $x(n) = s(n) + d(n)$, where n denotes time index. Further, we assume that $s(n)$ and $d(n)$ are statistically independent, and then whose variance is given by equation (4.1),

$$E\{x^2(n)\} = E\{s^2(n)\} + E\{d^2(n)\} \quad (4.1)$$

The input signal is either processed through WOLA or FFT Modulated filter bank, as expressed in details in section 2.4.1 in-order to analyze in time frequency domain. The analysis bank process the segmented input signal $x(\lambda R + n)$ by windowing and transforming into short-time spectrum $X(\lambda, k)$.

$$X(\lambda, k) = \sum_{n=0}^{N-1} w(n)x(\lambda R + n)e^{-j2\pi kn/N} \quad (4.2)$$

Where λ represents decimated time index, k is the frequency bin index, $k \in 0,1, \dots, N-1$, $w(n)$ is a window function and R denotes the block rate of number of samples to read in a frame.

4.1.2 Subband Power Estimation

To obtain the smoothed short time subband signal power $P_x(\lambda, k)$ equation (4.3), we perform squared magnitude of the output of the analysis bank and also apply a smoothing

factor α to the first order recursive equation. The smoothing constant value is given in between $\alpha = 0.90$ to 0.95 [21].

$$P_x(\lambda, k) = \alpha . P_x(\lambda - 1, k) + (1 - \alpha) . |X(\lambda, k)|^2 \quad (4.3)$$

4.1.3 Subband Noise Power Estimation

By using recursively smoothed periodograms, we estimate short time signal power $P_x(\lambda, k)$. By, taking this assumption we can estimate the noise power spectrum $P_n(\lambda, k)$, which tracks the minimum of the short time power estimate $P_x(\lambda, k)$ with a finite window length *L*.

$$P_x(\lambda, k) = omin . P_{min}(\lambda, k) \quad (4.4)$$

where $P_{min}(\lambda, k)$ is the minimum noise power and omin is a bias compensation factor. Due to computational complexity and delay the data window of length *L* is decomposed into *W* windows of length *M* i.e. *MW=L.* The window length must be large enough to bridge the broadest peak in the speech signal. It was experimentally proven that the window length of approximately 0.75 sec - 1.5 sec gives good results.

The minimum power of *M* samples is determined by sample-wise comparison within the window and the minimum values are updated into a buffer and the search for next minimum begins until last subband power. The obtained minimum noise power $P_{min}(\lambda, k)$ equation (4.5) is now compared with actual subband power $P_x(\lambda, k)$. If the actual subband power $P_x(\lambda, k)$ is smaller than the estimated noise power $P_n(\lambda, k)$, the minimum noise power is updated immediately [21].

$$P_{min}(\lambda + 1, k) = \min(P_x(\lambda, k), P_{min}(\lambda, k) \quad (4.5)$$

4.1.4 Compute SNR

The SNR is computed on the basis of estimated noise power $P_{min}(\lambda, k)$

$$SNR_x(\lambda, k) = 10 . \log(\frac{P_x(\lambda,k) - \min(P_n(\lambda,k), P_x(\lambda,k)}{P_n(\lambda,k)}) \quad (4.6)$$

We calculate $SNR_x(\lambda, k)$ because it forms the basis in deciding the oversubtraction factor. If we get high SNR value then the oversubtraction factor is less and if the *SNR* value is low we subtract with high value. The Berouti et. al. [22], have clearly explained about the relationship between the subband SNR_x and oversubtraction factor $osub$. By the proper selection of oversubtraction factor we can eliminate the residual noise which in fact improves the quality of speech by suppress the low energy phonemes [21].

$$osub(\lambda,k)=\begin{cases}1 & SNR(\lambda,k)\geq 20\\ 4-\frac{3}{20}SNR(\lambda,k), & -5\leq SNR(\lambda,k)\leq 20\\ 4 & SNR(\lambda,k)<-5\end{cases} \quad (4.7)$$

4.1.5 Subtraction Rule

The magnitude square of the input spectra is smoothed by first order recursive network ($\gamma \lesssim 0.9$) . The short time signal power can be written as,

$$\overline{|X(\lambda,k)|^2}=\gamma.\overline{|X(\lambda,k)|^2}+(1-\gamma).|X(\lambda,k)|^2 \quad (4.8)$$

The amount of subtraction is controlled by oversubtraction factor $osub(\lambda,k)$ and limitation to maximum subtraction by a spectral floor constant $subf(0.001\leq subf\leq 0.005)$.The spectral magnitudes are subtracted based on the following principle, given below [21].

$$|Y(\lambda,k)|=\begin{cases}\sqrt{subf.P_n(\lambda,k)}, & if \quad |X(\lambda,k)|.Q(\lambda,k)\leq\sqrt{subf.P_n(\lambda,k)}\\ |X(\lambda,k)|.Q(\lambda,k), & else\end{cases}$$

$$\text{where } Q(\lambda,k)=\left(1-\sqrt{osub(\lambda,k)\frac{P_n(\lambda,k)}{|X(\lambda,k)|^2}}\right) \quad (4.9)$$

After the subtraction part is done we add back the phase of the noisy speech spectrum to the output of the magnitude spectrum $|Y(\lambda,k)|$.It is then further processed through WOLA synthesis bank. The synthesis bank transforms the magnitude spectrum into the time domain signal, by applying IDFT and the output of the IDFT is circularly rotated and weighted by a window function which is then overlap added into the FIFO buffer to obtain an enhanced speech signal $y(t)$.

4.1.6 Description of parameters

The following parameters are important in implementing the spectral subtraction to get an enhanced speech signal.

4.1.6.1 Smoothing constant *(α):* The smoothing constant α is used in equation (4.3) to obtain recursively smoothed periodograms. The optimal choice of the smoothing is very important. If the estimated spectrum is smoothed too much then the peaks of the speech becomes broader and the small notches in the speech gets eliminated, which leads to inaccurate estimation of noise levels and valleys of the power in Figure 4.2 will not pronounced enough. The smoothing constant is set between α=0.9…0.95 [23].

4.1.6.2 Bias compensation factor (omin): This is the parameter used as bias compensation of minimum noise estimate. It sets the noise floor for the noisy speech signal whose values *omin*=0.99, to have smoothness in the reconstructed noise suppressed signal as shown in Figure 4.2 [23]. This value is concluded from set of random experiments conducted by varying value from 0.7 to 1.00 in achieving good perceptual sound listening quality

4.1.6.3 Window for minimum search (D): To get an effective noise power estimate choosing an appropriate window length is very important. It should be large enough to bridge any peaks of speech activity and short enough to follow non-stationary noise variations. The window length of 0.8 sec - 1.4 sec has proven to give good results [23].

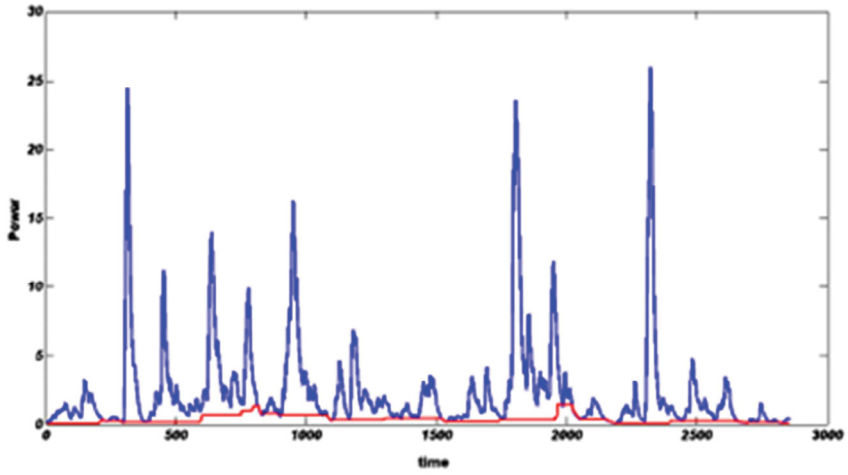

Figure 4.2 Estimate of smoothed power signal and the estimate of noise floor

4.1.6.4 Oversubtraction (*osub*) and spectral floor constant (*subf*):

The oversubtraction factor '*osub*' subtracts the estimate of the noise spectrum from noisy speech spectrum. After the subtraction there remain peaks in the spectrum. By using *osub* > 1, we can reduce the amplitude of the peaks and in some case it also eliminates them. By doing this, there remains a deep valley in spectrum surrounding the peaks. To avoid this, a spectral floor has been introduced. When we put *subf* > 0 then there is no more long deep valley between the peaks compared to *subf* = 0 and also masks the remaining spectral peaks by assigning a suitable spectral components.

4.2 Implementation in Matlab

The spectral subtraction algorithm is implemented based on the minimum statistics approach by Martin [21]. The SS algorithm is processed in spectral domain, using weighted overlap-add analysis and synthesis filter bank as shown in Figure 4.3. A 20ms hamming window with 50% overlap is used for the analysis. In the WOLA-FB, analysis bank processes the segmented time signal *x(n)* by applying the windowing function *w(n)* and then a FFT operation is used to transform into short-time subband signal. The output of the analysis bank is fed to the noise estimation block which estimates the short-time noise power. The estimated short-time noise power is then subtracted from short-time subband signal power based on the factors oversubtraction '*osub*' and spectral floor constant '*subf*'. These are the parameters which control the amount of subtraction on short-time subband signal power. In addition to the oversubtraction factor and the spectral floor constant, additional parameters are the smoothing factor (α), the window length for minimum search (*M*) and the bias compensation factor (*omin*) see section 5.2.6. The parameter values described above whose optimum values after several experiments we finalized to this thesis are: α=0.9 , *M*=200, *omin*=0.99. Further, after the subtraction part is done the output is added with phase component and then inverse FFT to give time domain signal. Finally, the enhanced speech output is produced by using the weighted overlap-add method. In the similar way by using the parameters defined for FFTMod-FB which are described in section 3.2.2 are used in implementing the MinSSS algorithm using the FFTMod-FB whose results are shown in chater-5.

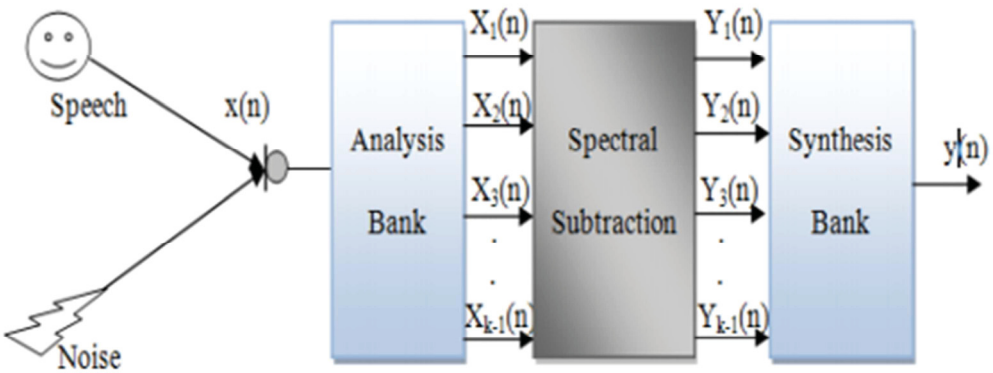

Figure 4.3 Spectral Subtraction using Filter Bank

Parameters	Values (Optimum)
minimum search (*M*)	200
bias compensation factor (*omin*)	0.99
smoothing factor (α)	0.9
oversubtraction factor (*osub*)	0.6

TABLE 4.1 OPTIMUM VALUES CHOSEN FOR MINSSS ALGORTHIM

CHAPTER-5

5. SIMULATION RESULTS AND ANALYSIS

In this thesis, we use ITU-T P.50 male and ITU-T P.50 female speech signal at the sampling frequency of 8 kHz as clean speech signal. ITU-T P.50 are the artificial voices that are used as test signals in telecommunication systems. The use of recommended artificial voices instead of real speech is the convenient way for the effective validation of the system. These ITU-T P.50 voices include 16 recorded sentences in each of 20 languages and are developed by some ITU members. Both signals are corrupted with the selected signal which is the combination of male and female voices, its phrase is "It's easy to tell the depth of a well⋯ kick the ball straight and follow through ⋯ blue the sheet to the dark blue background ⋯ a part of tea helps to past the evening" duration of 0.11 seconds or 182824 samples of data. The Speech signal is combined with the Gaussian Noise (GN), Car Noise (CAN), Factory Noise (FN), Wind Noise (WN) and Cafeteria Noise (CN) at 0 dB, 5 dB and 10 dB SNR for testing the system. This thesis makes use of the objective measures SNR improvement and PESQ for evaluation of the enhancement systems. Apart from these measures, there are other evaluation metrics.

5.1 Evaluation Parameters

Speech quality assessment is the basic step used to verify the robustness of any speech enhancement system. This assessment is based on different kinds of quality measures. Speech quality measures may classified into Subjective measure and Objective measure. Here we limit to objective quality parameters since subjective assessment involves the coordination of human subjects, they are typically associated with being time consuming, expensive and can require consideration of ethical issues.

The objective measure assess the quality of the output speech by comparing it with a reference sample, these measures can be classified into time, frequency or perceptual. Time domain objective measures tend to correlate least, eg: Signal to Noise Ratio (SNR). Frequency domain measures overcome the problem of time shift sensitivity of time domain measures. These are derived from zero phase representation such as Linear Prediction (LP) parameters or the frequency magnitude spectrum. Perceptual domain measures tend to be correlated with subjective data. These methods psycho-acoustic knowledge such as wrapped frequency scales and masking effects to estimate more perceptually meaningful distortion

scores. Some examples of such measures are Bark Spectral Distortion and Perceptual Evaluation of Speech Quality (PESQ).

5.1.1 Signal-to-Noise-Ratio (SNR)

SNR is a performance measure used in several applications, since this is essentially a sample to sample comparison, its use is most appropriate in situations where the intention is to reproduce the original speech signal exactly. The expression expressed in dB is

$$SNR = 10log_{10} \frac{\sum_{k=0}^{N-1} s(k)^2}{\sum_{k=0}^{N-1} [s(k)-y(k)]^2} \quad (5.1)$$

An SNR Improvement can be calculated by finding SNR at the input of the enhancement system and at output of system, then finally subtracting input SNR from output SNR i.e. *SNR-I = SNR_Out – SNR_In.* This SNR Improvement parameter indicates that system has improved SNR in the noisy speech signal by passing through a system.

5.1.2 Perceptual Evaluation of Speech Quality (PESQ).

This PESQ measure is described in ITU-T Recommendation P.862, the validation of PESQ included a number of experiments that specifically tested its performance over combinations of factors such as filtering, variable delay, coding distortions, and channel errors. PESQ compares an original clean speech signal *s(t)* with degraded signal y(t) that is the result of passing *s(t)* through communication systems, or with an enhanced signal $\hat{s}(t)$ calculated by the enhancement system. The output of PESQ is a perceived quality that would be given to *y(t)* or $\hat{s}(t)$ by subjects in a subjective listening test. The PESQ scores lies in the range 0 to 4.5 of MOS scale.

MOS Scores	Conclusion
0 - 0.5	Very bad
0.5 - 1.5	Bad
1.5 - 3	Good
3 - 4.5	Excellent

Table 5.1 MOS (Mean Opinion Score) Scores [25]

5.1.3 Total Noise Level Reduction (TNLR)

The total noise level reduction measure, or TNLR, relates to the capability of the NR method to attenuate the background noise level measured during both speech activity and speech pauses. As the number of frames in the speech pause class during speech activity is typically relatively small compared to the number of frames during long speech pauses, TNLR mainly measures the capability of an NR to reduce noise during long speech pauses.

This TNLR estimates the overall level of noise reduction experienced both during speech and speech pauses. The TNLR measure is calculated as follows: For each background noise condition j and for each speech sample i, construct a noisy input signal d_{ij} as follows:

$$d_{ij}(n) = \beta_{ij}\, n_j(n) + s_i(n), \tag{5.2}$$

whereβ_{ij} depends on the *SNR* condition according to the procedure described above. The noise suppressed output speech signal can be denoted as:

$$y_{ij} = NR\ (d_{ij}) \tag{5.3}$$

In this way, the total noise level reduction measure can be expressed as:

$$TNLR_{ij} = \frac{1}{K_{pse}} 10 \cdot \sum_{m=1}^{K_{pse}} \left[\mathrm{Log}\left\{ \max\left(\xi, \sum_{q} d_{ij}^2(m,q) \right) \right\} - \mathrm{Log}\left\{ \max\left(\xi, \sum_{q} y_{ij}^2(m,q) \right) \right\} \right] \tag{5.4}$$

where: K_{pse} is the total number of noise frames during speech pauses (frame power of clean speech signal <sp_lvl + th_nh) with active noise (frame power of noisy input signal >cft_lvl; $\xi > 0$ is a constant that should be set at 8.10^{-8}; and cft_lvl is a comfort noise level constant that will default to –48 dB

$$TNLR_j = \frac{1}{\mathrm{I}} \sum_{i=1}^{I} TNLR_{ij} \tag{5.5}$$

$$TNLR = \frac{1}{\mathrm{J}} \sum_{j=1}^{J} TNLR_j \tag{5.6}$$

The summation with respect to index *q* is carried out in noise frames of 80 samples. The index, *q*, relates to noise frames with frame power less than the higher bound for speech pause class. Furthermore, it is informative to record separately the noise type specific TNLR measures, or $TNLR_j$, for each background noise condition *j*.

5.1.4 SNRI-to-NPLR difference (DSN)

DSN, comprising a comparison of the SNRI and NPLR measures, is therefore proposed as a measure to acquire an indication of possible speech attenuation or speech amplification produced by the tested NR method and is formulated as:

$$DSN = SNRI - NPLR \tag{5.7}$$

Typically the DSN quantity should get values close to zero. If the NPLR parameter assumes clearly higher values than SNRI, making DSN clearly negative, the NR solution turns out to produce speech level attenuation. On the other hand, if DSN becomes clearly positive, it indicates the speech signal has been amplified, which contributes to the SNR improvement partially or wholly if there is no reduction in noise level. This measurement (*DSN*) is computed to reveal speech attenuation or undesired speech amplification caused by an NR solution.

5.2. WOLA-FB Minimum Statistics Spectral Subtraction

In this section we show the performance of MinSSS algorithm using WOLA-FB for various types of noises at different SNR dB levels. For all types of noises the amount of noise removed is high especially at low SNR inputs (0dB, 5dB and 10dB), i.e. SNR-I is higher when compared to SNR-I at higher SNR inputs of 15dB and 20dB. But at lower SNR inputs the quality of speech in reconstructed signal is less when compared to higher SNR inputs level i.e. in Table 5.1 the PESQ-I columns shows that the amount of PESQ score is improved after cancelling the noise in the noisy speech signal. In the DSN column of table, the negative values signifies the level of speech is reduced and where as the positive values represents the level of speech level is increased when compared to original speech signal.

Type of noise	SNR In	SNR Out	SNR-I	PESQ In	PESQ Out	PESQ-I	TLNR	DSN
White	0	10.6193	10.6193	1.301	1.649	0.348	14.1223	-2.5532
	5	15.4503	10.4503	1.565	2. 164	0.599	13.6161	-1.0153
	10	18.7305	8.7305	1.891	2. 577	0.686	13.2774	-0.1521
	15	21.6505	6.6505	2.286	2.800	0.514	12.5743	0.5333
	20	23.4827	3.4827	2.624	2.989	0.365	12.0244	1.07679
Pink	0	8.5161	8.5161	1.322	1.661	0.339	13.9972	-4.6883
	5	13.5692	8.5692	1.618	2.131	0.513	13.7780	-2.8102
	10	18.6521	8.6521	1.968	2.509	0.541	13.4574	-0.9452
	15	21.3048	6.3048	2.361	2.779	0.418	12.6645	0.1764
	20	24.6531	4.6531	2.695	3.007	0.312	12.8007	0.8237
Car	0	10.2276	10.2279	1.849	2.267	0.418	13.4613	-2.2610
	5	13.3831	8.3831	2.248	2.46	0.212	12.8517	-1.4234
	10	18.5534	8.5534	2.523	2.777	0.254	12.9860	0.2619
	15	20.2621	5.2621	2.822	3.121	0.299	12.1714	1.1965
	20	23.5831	3.5834	3.121	3.293	0.172	13.0290	1.8116
Engine	0	12.9726	12.9726	2.469	2.853	0.384	14.0147	0.1086
	5	16.9661	11.9661	2.762	3.148	0.386	14.2527	1.4389
	10	19.8179	9.8179	3.037	3.315	0.278	13.8025	1.7532
	15	21.5969	6.5969	3.301	3.451	0.15	13.3834	2.0985
	20	24.8854	4.8854	3.617	3.513	-0.104	14.1116	2.3693
Factory	0	9.26793	9.2679	1.863	2.256	0.393	12.4973	-1.6054
	5	14.8418	9.8418	2.263	2.565	0.302	12.3340	0.1349
	10	17.7842	7.7842	2.586	2.85	0.264	11.9339	0.8029
	15	20.4452	5.4452	2.88	3.123	0.243	11.5203	1.5045
	20	22.6884	2.6884	3.156	3.34	0.184	12.9158	1.9328

Table 5.2 SNR-I, PESQ-I, TLNR, DSN for all noises at different dB levels using WOLA-FB and MinSSS

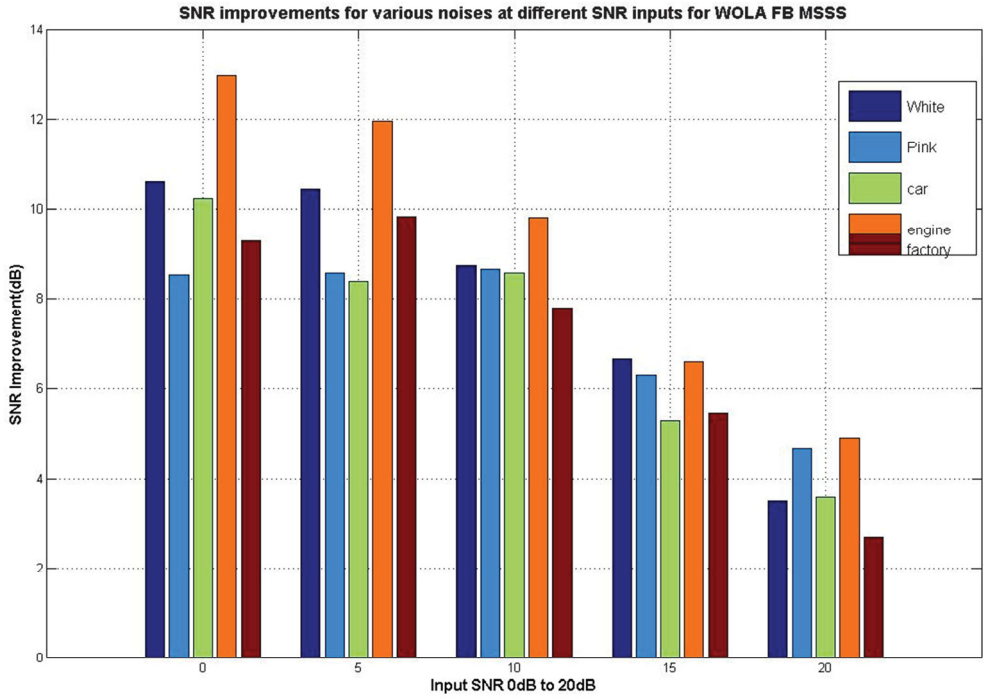

Figure 5.1 SNR-I for different noises at different dB levels using WOLA-FB and MinSSS

5.3 FFT Modulated Filter Bank Minimum Statistics Spectral Subtraction

This section shows the performance of MinSSS algorithm using FFT-Modulated FB for various noises at different SNR dB levels. For all noises the amount of noise cancelled is high at low SNRs i.e. at 0dB and 10dB SNR input when compared at higher SNR inputs, i.e. which resembles that the SNR improvement is higher when compared to at higher SNR inputs at 15 and 20dB. But at lower SNR inputs the quality of speech in reconstructed signal is less when compared to higher SNR inputs levels i.e. in table the PESQ-I columns shows that how much PESQ score is improved after cancelling the noise in the noisy speech signal. In the DSN column of table, the negative values signifies that the level of speech is reduced and where as the positive values represents the level of speech level is increased when compared to original pure speech signal.

Type Of noise	SNR In	SNR Out	SNR-I	PESQ In	PESQ Out	PESQ-I	TLNR	DSN
White	0	10.8345	10.8345	1.301	1.643	0.342	20.5801	-9.3241
	5	15.8237	10.8237	1.565	2.221	0.656	20.3049	-7.4600
	10	19.6245	9.6242	1.891	2.629	0.738	19.8244	-6.4575
	15	22.4657	7.4652	2.286	2.884	0.598	19.3098	-5.8180
	20	25.0087	5.0084	2.624	3.103	0.479	18.7615	-5.3659
Pink	0	8.94707	8.9470	1.322	1.67	0.348	20.7299	-11.215
	5	14.2163	9.2161	1.618	2.189	0.571	20.4185	-9.2626
	10	18.8146	8.8146	1.968	2.562	0.594	19.9923	-7.4546
	15	22.0457	7.0450	2.361	2.849	0.488	19.6333	-6.3508
	20	25.8329	5.8328	2.695	3.088	0.393	19.2851	-5.5246
Car	0	11.3757	11.3756	1.849	2.259	0.41	19.8205	-8.8368
	5	14.2919	9.2919	2.248	2.626	0.378	19.6093	-7.7126
	10	19.1671	9.1671	2.523	2.893	0.37	19.4041	-6.1125
	15	21.4680	6.4680	2.822	3.243	0.421	18.6153	-5.1226
	20	24.8790	4.8790	3.121	3.493	0.372	20.0037	-4.4668
Engine	0	13.7113	13.7113	2.469	2.993	0.524	21.2067	-5.9494
	5	18.2650	13.2650	2.762	3.282	0.52	20.8108	-4.6262
	10	20.589	10.5889	3.037	3.514	0.477	20.2956	-4.3631
	15	22.2818	7.2818	3.301	3.685	0.384	19.2420	-4.1312
	20	25.3975	5.3975	3.617	3.817	0.2	20.2491	-3.9124
factory	0	11.7531	11.7531	1.863	2.336	0.473	19.1423	-7.7033
	5	15.2808	10.2808	2.263	2.638	0.375	19.0279	-6.1446
	10	17.8648	7.8647	2.586	2.923	0.337	18.4184	-5.7389
	15	21.2872	6.2872	2.88	3.222	0.342	17.9745	-4.8487
	20	24.4356	4.4356	3.156	3.509	0.353	19.3996	-4.392

Table 5.3 SNR-I, PESQ-I, TLNR, DSN for all noises at different dB levels using FFT-modulated FB and MinSSS

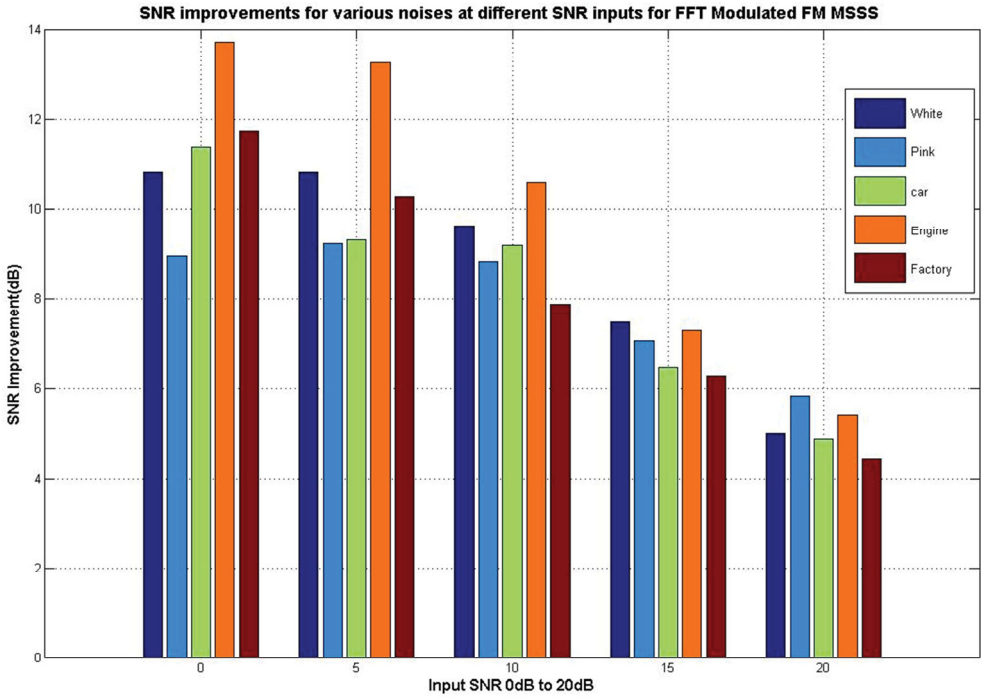

Figure 5.2 SNR-I for different noises at different dB levels using FFT-Modulated FB and MinSSS

5.4 Comparison of WOLA-FB MinSSS and FFTMod-FB MinSSS

From the Table 5.3, it clearly shows that the performance of MinSSS algorithm using FFTMod-FB outperforms when using with WOLA-FB i.e. both in terms of SNR-I and PESQ-I. The SNR-I represents amount of noise levels decreased in reconstructed signal and whereas PESQ-I indicates the score of quality of speech increased in reconstructed speech signalwhen compared to original noisy signal. From Figure 5.3 to Figure 5.12 compares the performances of WOLA-FB-MinSSS and FFTMod-FB-MinSSS in terms of SNR-I and PESQ-I for all type of noises at different levels of input SNR's (dB).

Type of Noise	SNR Inputs	SNR Improvement(SNR-I)		PESQ Improvement (PESQ-I)	
		WOLA-FB	FFT_Modulated FB	WOLA-FB	FFT-Modulated-FB
White	0	10.6193	10.8345	0.348	0.342
	5	10.4503	10.8237	0.599	0.656
	10	8.7305	9.6242	0.686	0.738
	15	6.6505	7.4652	0.514	0.598
	20	3.4827	5.0084	0.365	0.479
Pink	0	8.5161	8.9470	0.339	0.348
	5	8.5692	9.2161	0.513	0.571
	10	8.6521	8.8146	0.541	0.594
	15	6.3048	7.0450	0.418	0.488
	20	4.6531	5.8328	0.312	0.393
Car	0	10.2276	11.3756	0.418	0.41
	5	8.3831	9.2919	0.212	0.378
	10	8.5534	9.1671	0.254	0.37
	15	5.2621	6.4680	0.299	0.421
	20	3.5831	4.8790	0.172	0.372
Engine	0	12.9726	13.7113	0.384	0.524
	5	11.9661	13.2650	0.386	0.52
	10	9.8179	10.5889	0.278	0.477
	15	6.5969	7.2818	0.15	0.384
	20	4.8854	5.3975	-0.104	0.2
Factory	0	9.2679	11.7531	0.393	0.473
	5	9.8418	10.2808	0.302	0.375
	10	7.7842	7.8647	0.264	0.337
	15	5.4452	6.2872	0.243	0.342
	20	2.6884	4.4356	0.184	0.353

Table 5.4 Comparision of SNR-i and PESQ-I between WOLA and FFT-Modulated FB's for different noises at various levels

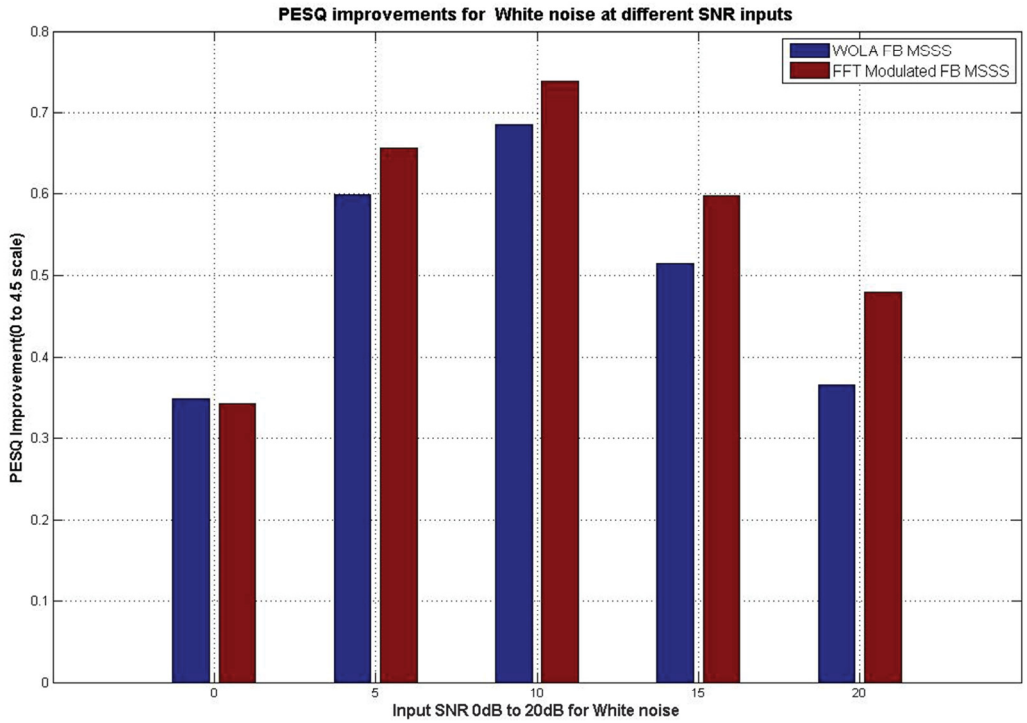

Figure 5.3 PESQ-I for white noise at different dB levels for WOLA-MinSSSand FFTMod-MinSSS

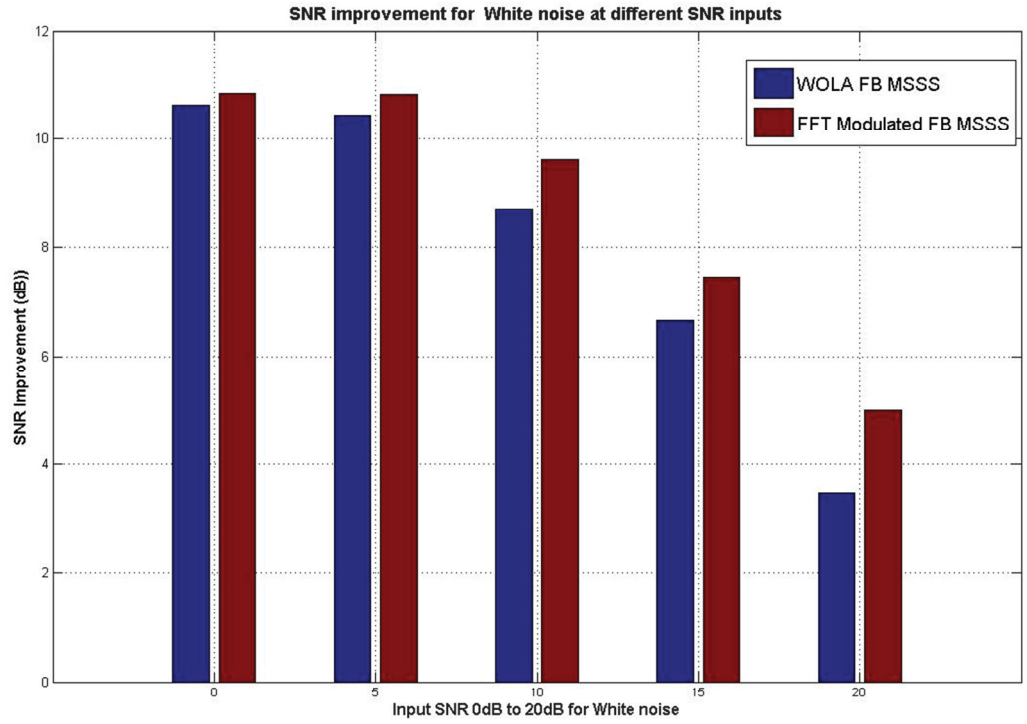

Figure 5.4 SNR-I for white noise at different dB levels for WOLA-MinSSS and FFTMod-MinSSS

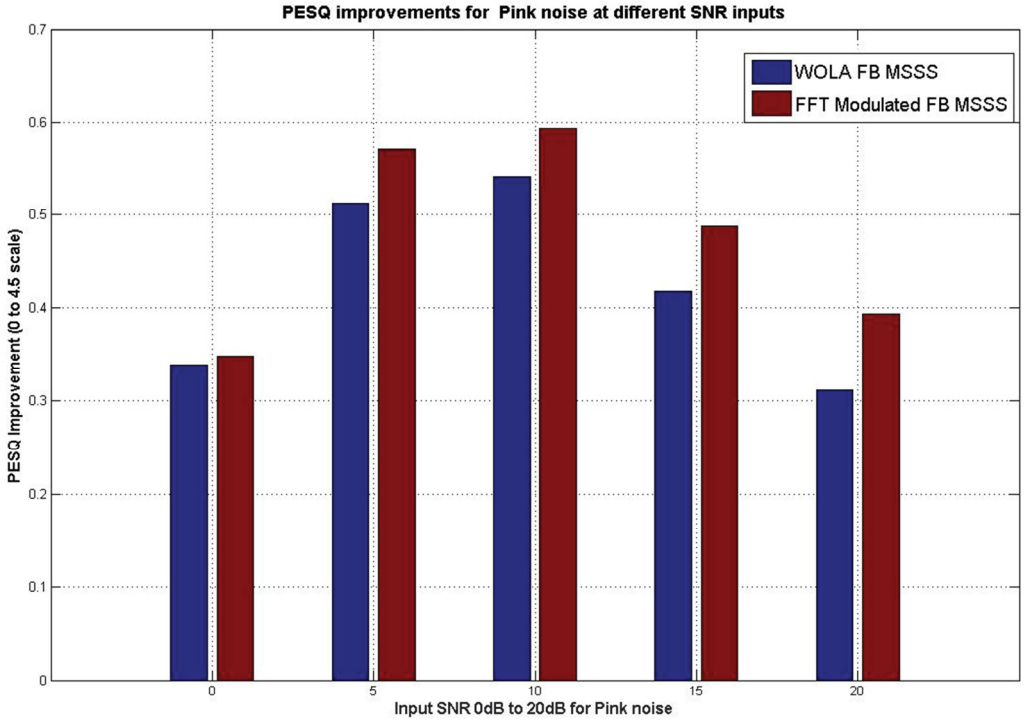

Figure 5.5 PESQ-I for Pink noise at different dB levels for WOLA-MinSSS and FFTMod-MinSSS

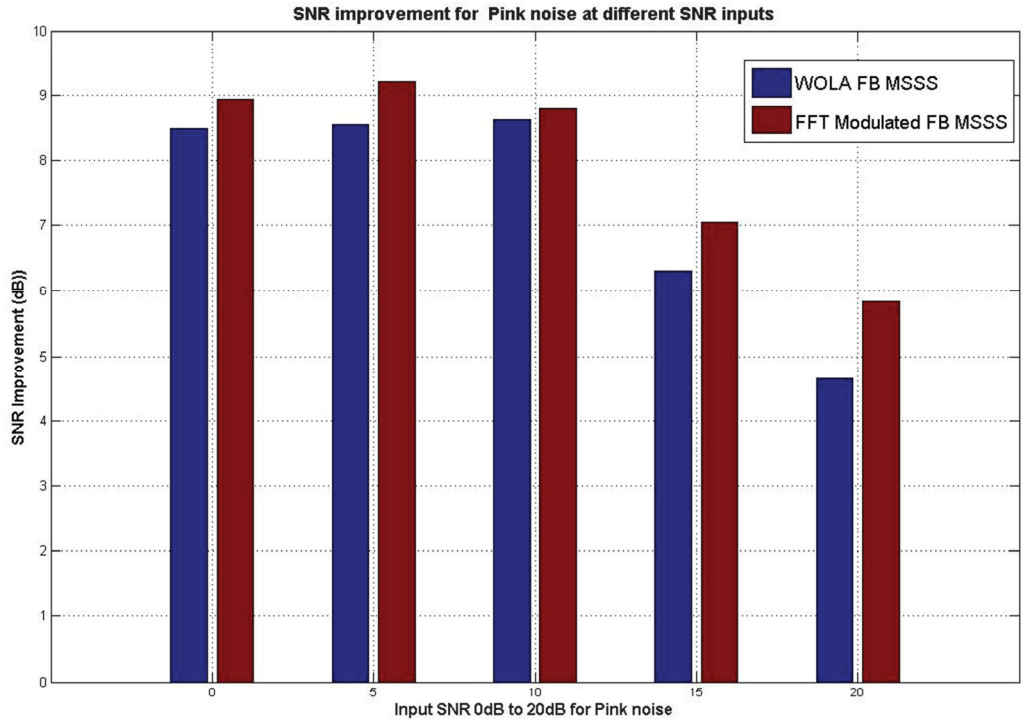

Figure 5.6 SNR-I for Pink noise at various dB levels for WOLA-MinSSS and FFTMod-MinSSSS

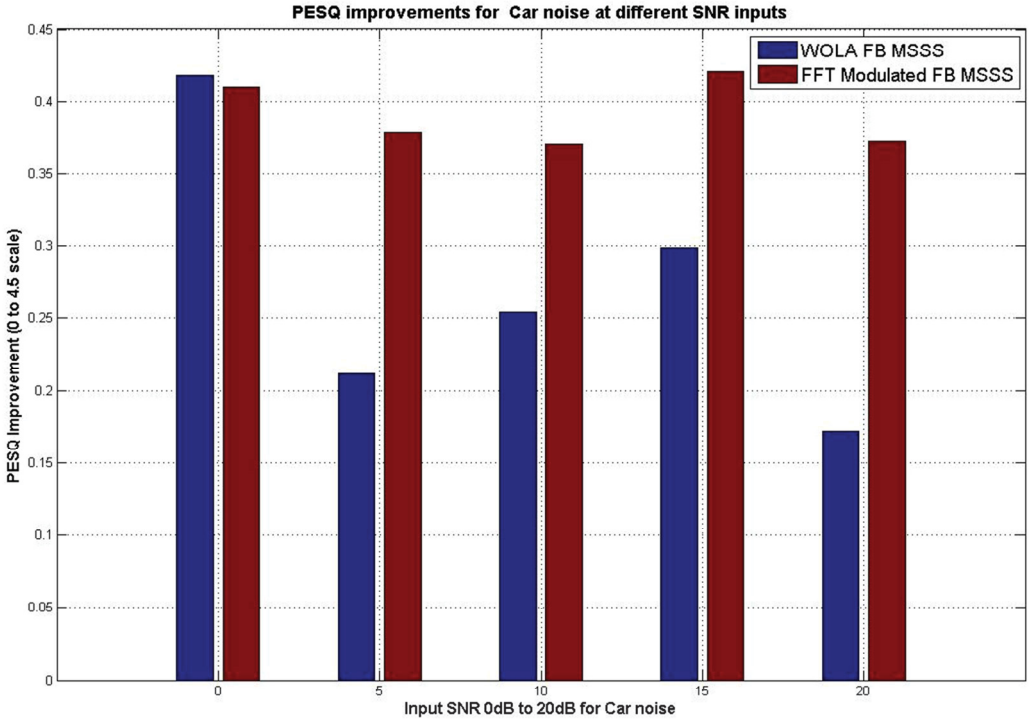

Figure 5.7 PESQ-I for car noise at various dB levels for WOLA-MinSSS and FFTMod-MinSSS

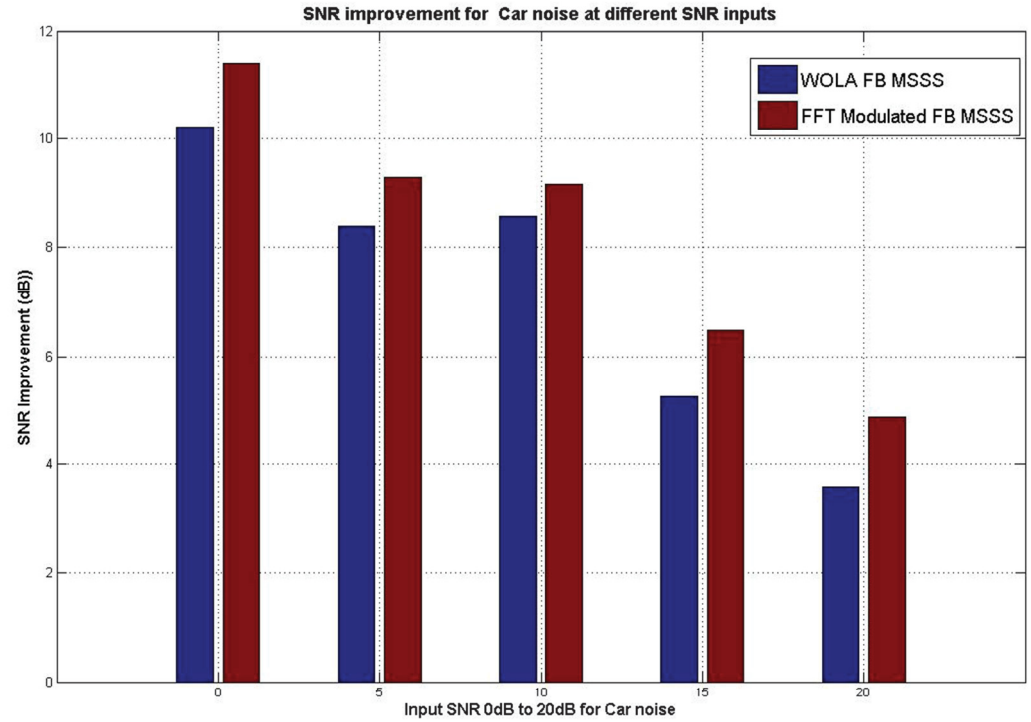

Figure 5.8 SNR-I for car noise at various dB levels for WOLA-MinSSS and FFTMod-MinSSS

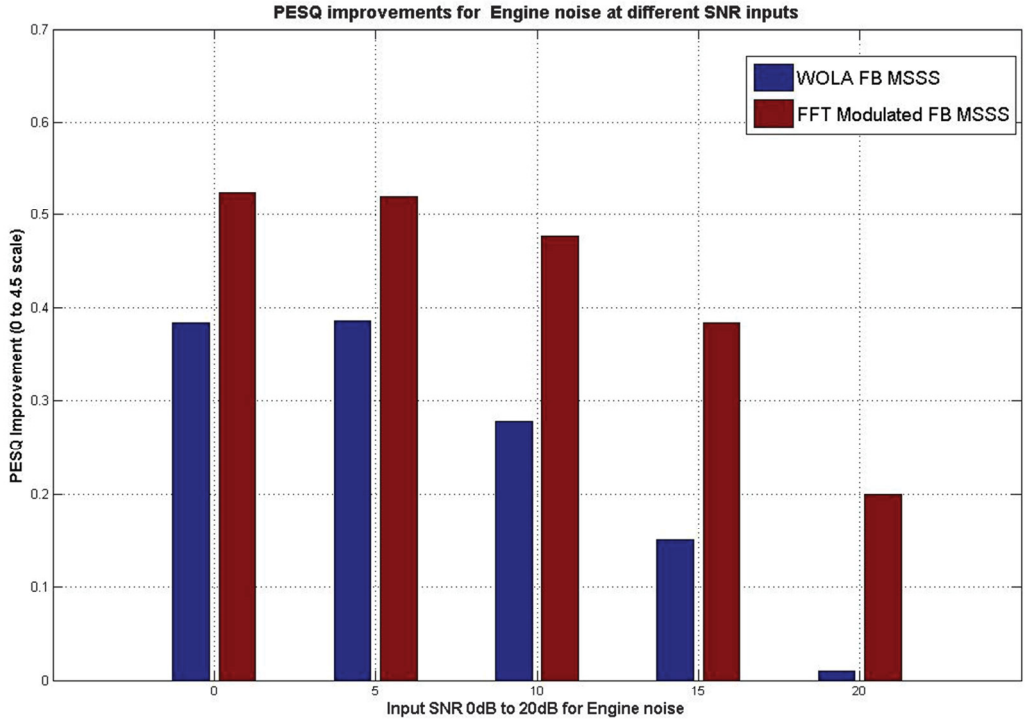

Figure 5.9 PESQ-I for Engine noise at various dB levels for WOLA-MinSSS and FFTMod-MinSSS

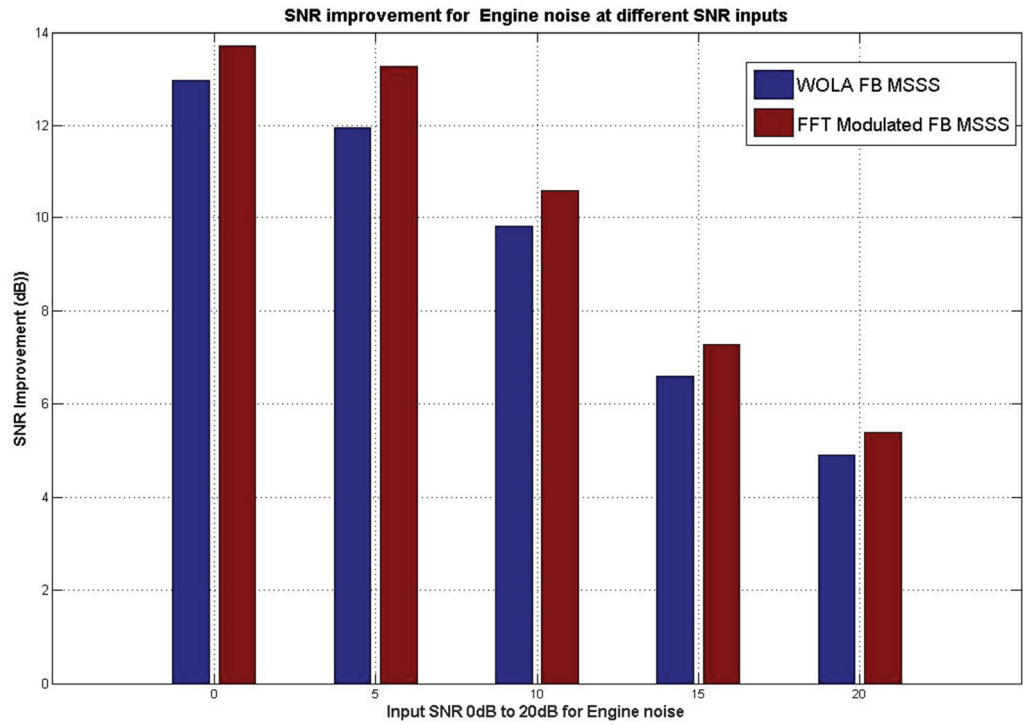

Figure 5.10 SNR-I for Engine noise at various dB levels for WOLA-MinSSS and FFTMod-MinSSS

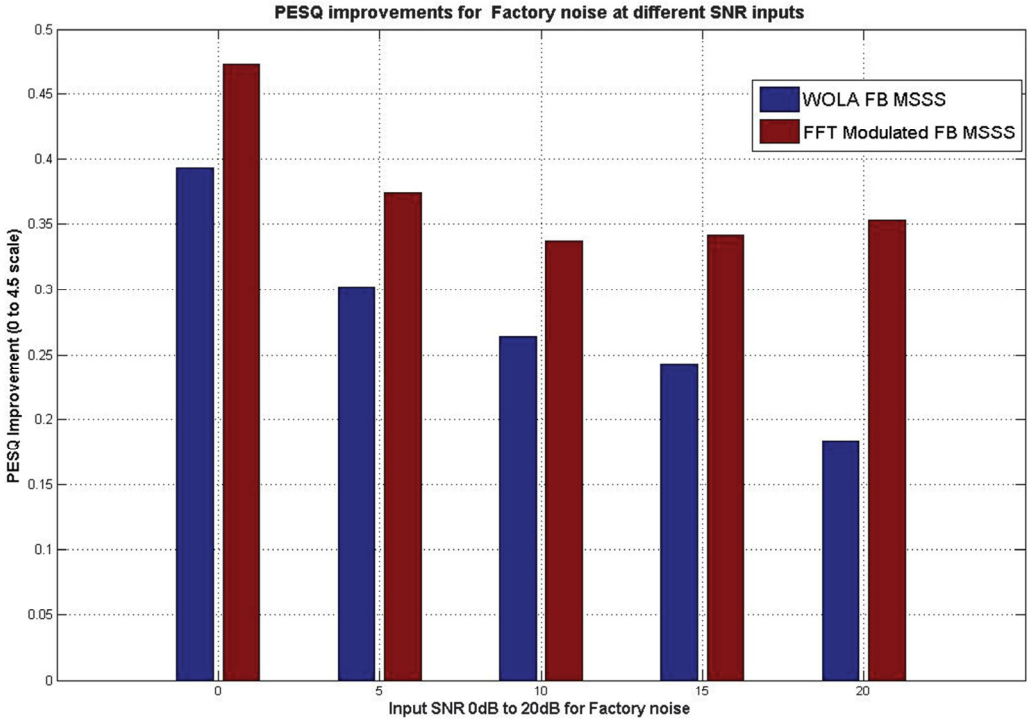

Figure 5.11 PESQ-I for Factory at various dB levels for WOLA-MinSSS and FFTMod-MinSSS

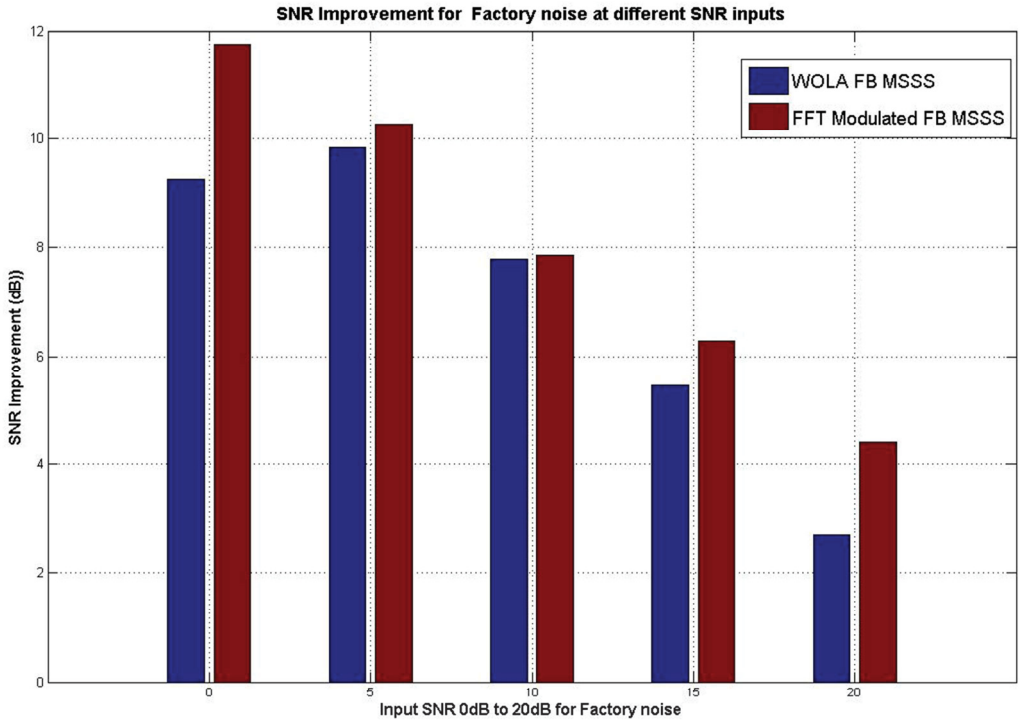

Figure 5.12 SNR-I for Factory noise at various dB levels for WOLA-MinSSS and FFTMod-MinSSS

CHAPTER-6

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

This thesis aims at evaluating the performance of Minimum Statistics of Spectral Subtraction using different filter bank approaches such as WOLA and FFT-Modulated based filter bank. Our implementation of the MinSSS method affords a great reduction in the background noise with very little effect on the intelligibility of the speech. The performance is measured using objective metrics such as SNR and PESQ. To evaluate the systems [WOLA-FB MinSSS and FFTMod-FB MinSSS] the clean speech signal is corrupted with five different types of noise signals which occur frequently during natural conversation in anechoic environment only.The systems are designed offline by simulating in Matlab-2011. The simulation results show that MinSSS based on FFT Modulated FB outperforms the MinSSS based on WOLA FB in terms of PESQ and SNR objective quality measures for five different kinds of noise signals corrupted at various noise energy levels. But the algorithm especially proved to be robust in anechoic environment, even down to SNRs as low as 0 dB. This thesis gives brief overview in choosing filter bank parameters such as number of frequency bins, ratio of overlap between the frames and frequency resolution, length of FB which plays vital role in designing lossless reconstruction filter bank. Also taken much careful choosing optimal parameters values in implementing MinSSS such as smoothing constant (α=0.90), bias compensation factor (*omin*=0.99), oversubtraction (*osub*=0.6) and spectral floor constant (*subf*), Window for minimum search (M=200).

6.2 Future work:

The implemented MinSSS algorithm eliminates the need for a speech activity detector by exploiting the short time characteristics of speech signals. But still there was difficulty of removing all the noise components from noisy speech without introducing musical noises or distortions, hence in this regard further research can be conducted to increase the accuracy of noise estimation and also the more adjustment needed to improve the trade-off between the smoothing constant and the window length for minimum search. As we have implemented it in offline mode so there is great scope to implement it on real time i.e. on specific kind of processor.

CHAPTER-7

7.BIBLIOGRAPHY

[1] R. A. Monzingo and T. W. Miller, Introduction to Adaptive Arrays, New York: John Wiley and Sons, 1980.

[2] Gray W Elko, Microphone Arrays: Signal Processing Techniques and Applications, chapter 17. Future Directions for Microphone Arrays, Springer-Verlag, 2001.

[3] Boll, S. F , "Suppression of acoustic noise in speech using spectral subtraction," Acoustics, Speech and Signal Processing, IEEE Transactions on , vol.27, no.2, pp. 113- 120, Apr 1979.

[4] J. Deller Jr., J. Hansen and J. Proakis, "Discrete-Time Processing of Speech Signals", *NY: IEEE Press*, 2000.

[5] H. Levitt, "Noise reduction in hearing aids: An overview", *Journal of RehabilitationResearch and Development,* vol. 38, No. 1, January/February 2001.

[6] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 26, No. 3, pp. 197-210, June 1978.

[7] R. E. Crochiere and L. R. Rabiner, "Mulitirate digital signal processing", Prentice-Hall, 1983.

[8] J. J. Shynk, "Frequency Domain and Muliti-Rate Adaptive Filtering," *IEEE Signal Processing Magazine,* vol. 9, 1992.

[9] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE,* vol. 67, No. 12, pp. 221-239, Dec. 1979.

[10] N. Virag, "Speech enhancement based on masking properties of the human auditory system," Master thesis, Swiss Federal Institute of Technology, 1996.

[11] Y. Ephraim, "Statistical Model Based Speech Enhancement Systems", in Proc. IEEE, vol. 80, no. 10, pp. 1526-1555, October 1992.

[12] Y. Gong, "Speech Recognition in Noisy Environments: A survey", in Speech Communication, Elsevier Science B.V., Amsterdam, Netherlands, vol. 16, no.3, pp 261-291, April 1999.

[13] M. Sambur, "Adaptive noise canceling for speech signals," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 26, pp. 419-423, 1978.

[14] M. Schroeder, “Models of hearing,” *Proc. IEEE,* vol. 63, No. 9, pp. 1332-1350, Sept.1975.

[15] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol.27, pp. 113-120, Apr. 1979.

[16] D. Gabar, "Theory of communication," *J, IEE,* no. 93, pp. 429-457, Nov 1946.

[17] R. E. Crochiere, "A weighted overlap-add method of short-time fourier analysis/snthesis," in *IEEE Trans. Acoust. Speech and Sig. Proc*, 1980, 1980.

[18] R. Brennan and T. Schneider, "A flexible filterbank structure for extensive signal manipulations in digital hearning aids," *Proc. IEEE Int. Symp. Circuits and Systems,* pp. 569-572, 1988.

[19] G. D. Cain, A. Yardim and P. Henry, "Offset windowing for FIR fractional-sample delay," *Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing (ICASSP-95),* vol. 2, pp. 1276-1279, 9-12 May 1995.

[20] S. M. Ahadi, H. Sheikhzadeh, R. L. Brennan and G. H. Freeman, "A Weighted Overlap Add-based Front-end for Speech Recognition," Journal of Iranian Association of Electrical and Electronics Engineers, vol. 1, no. 2, pp. 15-20, 2004.

[21] R. Martin, “Spectral Subtraction Based on Minimum Statistics”, in Proc. EUSPICO’94, pp. 1181-1185, 1994.

[22] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of Speech Corrupted by Acoustic Noise”, in Proc. IEEE Conf. ASSP, pp. 208-211, April 1979.

[23] Jan Mark De Haan, “Filter bank design for digital speech signal processing”.

[24] Y. Ephraim, “Statistical-model-based speech enhancement systems,” *Proc. IEEE,* vol. 80, No.10, pp. 1526-1555, Oct.1992.

[25] Philipos C. Loizou, Speech Enhancement: Theory and Practice, CRC Taylor and Francis, 2007.

[26] R. Martin, “An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals”, in Proc. EU-ROSPEECH ‘93, pp. 1093-1096, Berlin, September 21-23, 1993.