TRANSFORM BASED MULTI-BAND SPEECH ENHANCEMENT ALGORITHMS

NAVNEET UPADHYAY

DEPARTMENT OF ELECTRICAL & ELECTRONICS ENGINEERING
BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
PILANI - 333031, INDIA
APRIL, 2013

Transform Based Multi-Band Speech Enhancement Algorithms

THESIS

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

by

NAVNEET UPADHYAY
Department of Electrical & Electronics Engineering

Under the Supervision of
Dr. Abhijit Karmakar

to the

Birla Institute of Technology & Science, Pilani
Pilani - 333031, India
April, 2013

CERTIFICATE

This is to certify that the thesis entitled "**Transform Based Multi-Band Speech Enhancement Algorithms**" being submitted by Mr. Navneet Upadhyay, ID No.: 2009PHXF035P, to the Department of Electrical & Electronics Engineering, Birla Institute of Technology & Science, Pilani for the award of the degree of Doctor of Philosophy is a bonafide record of research work carried out by him under my supervision. He has fulfilled all the requirements for submission of the thesis which has reached the requisite standard.

I hereby declare that the content of this thesis, in full or in parts, to the best of my knowledge, has not been submitted to any other educational Institute or University for the award of any other degree or diploma.

Date : April 30, 2013
Place : Pilani

(Dr. Abhijit Karmakar)
Associate Processor, AcSIR[!]
Principal Scientist,
CSIR - Central Electronics Engineering Research Institute
Pilani - 333031, India

[!] Academy of Scientific & Innovative Research

ABSTRACT

Index Terms — *single channel speech enhancement*; *adverse environment; additive noise*; *noise estimation*; *short-time Fourier transform*; *stationary wavelet packet filterbank*; *critical band rate scale*; *spectral subtractive-type algorithms*; *iterative processing*; *spectrogram*; *objective measure*; *subjective measure.*

The degradation of speech due to the presence of additive background noise causes severe problems in a variety of communication environments. The spectral subtraction method is a classical approach and is widely used for enhancement of degraded speech. The drawback of this classical approach is that it uses a fixed set of values for the subtraction parameters. Thus, its applications are limited only for the noises that degrade the speech signal uniformly. However, real-world noise is mostly non-stationary or colored in nature and a multi-band approach is found to be more efficient than the classical approach.

This thesis mainly addresses the problem of single channel speech enhancement, where the signal is derived from a single microphone, in adverse environment and proposes transform based multi-band speech enhancement algorithms. The aim of the proposed research work is to explore the transform based multi-band algorithms for the augmentation of the overall quality and intelligibility of the processed speech by suppressing the background noise as well as the remnant component of the noise. The algorithms proposed in the thesis are based on short-time Fourier transform and stationary wavelet packet transform and the noise reduction is performed on the transform coefficients by using adaptive noise estimation approach. The performance of the algorithms has been evaluated comprehensively and their comparative study has been done.

The thesis presents a detailed study of the spectral subtractive-type algorithms with a unified view of the single channel speech enhancement algorithms in the frequency domain, which provides the necessary algorithmic framework required for the development of the proposed transform based multi-band speech enhancement algorithms. The study includes the comparative analysis of the different forms of spectral subtractive-type algorithms such as the basic magnitude spectral subtraction algorithm, spectral over-subtraction algorithm, multi-band spectral subtraction algorithm, Wiener filtering, iterative spectral subtraction algorithm, and spectral subtraction based on perceptual properties in noisy environments.

The iterative processing based multi-band spectral subtraction algorithm proposed in the thesis aims to enhance the narrowband speech degraded by non-stationary noises. The speech is processed into four uniformly spaced continuous frequency bands and the spectral subtraction is performed independently on each band using band specific over-subtraction factors. This process is iterated a small number of times and the noise is estimated in each iteration. In this scheme, the output from the base enhancement stage is used as the input for next iteration process, where the additive noise changes its form into remnant noise and subsequently, this remnant noise is re-estimated in each iteration.

An improved multi-band spectral subtraction based on critical band rate sale of human auditory system is explored next. Here, the speech degraded by non-stationary noise is processed by splitting the complete spectrum into six non-uniformly spaced frequency bands in accordance to the critical band rate scale and spectral subtraction is applied independently in each band using band specific over-subtraction factors. The enhancement algorithm uses an adaptive approach to estimate the noise from each band without the need of explicit speech silence detection.

The thesis also proposes a perceptually motivated stationary wavelet packet transform based multi-band improved spectral over-subtraction algorithm for the enhancement the speech degraded in adverse environment. The perceptually motivated stationary wavelet packet filterbank is used to decompose the input noisy speech signal into seventeen non-uniform subbands and the speech is processed independently in each subband using improved spectral over-subtraction. The adaptive noise estimation approach used in this algorithm requires no voice activity detection and thus, it can update the noise estimate throughout the signal instead of being limited to estimating the noise in silence intervals. This allows a more accurate noise estimate and thereby, improves the quality of the enhanced speech further.

All the proposed algorithms are evaluated and compared with contemporary speech enhancement algorithms in terms of the quality and intelligibility of the enhanced speech. Various objective measures such as signal-to-noise ratio (SNR), segmental SNR (SegSNR), Itakura-Saito distortion (ISD), and perceptual evaluation of speech quality (PESQ) are performed on the test set, obtained by degrading three male utterances and a female utterance with seven different types of real-world noises and a computer generated stationary noise at different levels of SNRs, ranging from 0 to 15 dB. A subjective listening test based on mean opinion score (MOS) is also carried out along with spectrogram analysis. The results of the subjective tests are also compared with those of the objective measures. The strengths and weaknesses of various proposed algorithms are analyzed and compared.

ACKNOWLEDGEMENT

I would like to express my sincere thanks to my supervisor Dr. Abhijit Karmakar, for giving me opportunity to work in his research group, and for supervising the work carried out in this thesis. He has offered me many helpful suggestions during the course of this research work.

I would also like to thank my doctorial advisory committee (DAC) members, Prof. Vinod Kumar Chaubey, and Dr. Karunesh Kumar Gupta for providing me valuable suggestions, feedback and moral support throughout the course of this journey.

I would also like to take this opportunity to thank the listeners who have spared their valuable time for participating in the subjective listening tests carried out in this thesis work.

My wife, Mrs. Neelam Upadhyay, has been the most helpful during the writing of this thesis. Besides her continuous moral support and encouragement, she has also done proofreading (more than once!) of the manuscript.

Finally, I express my deepest and most sincere regards to my parents for their moral support, loving encouragement and understanding that they have shown during the whole duration of my doctoral studies. They have always encouraged me to fulfill my dreams.

Date : April 30, 2013 (Navneet Upadhyay)
Place : Pilani

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$s(n)$	Clean speech signal
$d(n)$	Noise signal
$y(n)$	Noisy speech signal
n	Discrete time index
N	Number of samples in the speech signal
ω	Discrete frequency index of the frame
α	Over-subtraction factor
β	Spectral floor parameter
α_i	Over-subtraction factor for i^{th} Band
δ_i	Additional band over-subtraction factor for i^{th} Band
f_s	Sampling frequency
$H(\omega)$	Frequency response of the spectral subtraction filter
$T(\omega)$	Masking threshold
$\mathbb{L}^2(\mathbb{R})$	Square-integral functions in $\mathbb{R}$
$\mathbb{R}$	Set of real numbers
$\psi(t)$	Mother wavelet
$\mathbb{Z}$	Set of integers
$\lambda(\omega,k)$	Smoothing factor
T	Center offset of the transition curve
a	Parameter in sigmoid function
K	Total number of bands
$z(f)$	Critical band rate scale in Bark
J	Maximum number of levels of SWPT

LIST OF ACRONYMS

ASR	Automatic Speech Recognition
AWGN	Additive White Gaussian Noise
ACR	Absolute Category Rating
API-MBSS	Auditory Perception based Improved Multi-Band Spectral Subtraction
BSS	Basic Spectral Subtraction
BS-WPD	Bark Scaled Wavelet Packet Decomposition
CWT	Continuous Wavelet Transform
CB	Critical Band
CBW	Critical Bandwidth
DWT	Discrete Wavelet Transform
dB	deciBell
Db	Daubechies
ER	Enhancement Rate
FWR	Full-Wave Rectification
FIR	Finite Impulse Response
FFT	Fast Fourier Transform
HWR	Half-Wave Rectification
HPF	High pass Filter
HAS	Human Auditory System
HMM	Hidden Markov Model
IP-MBSS	Iterative Processing based Multi-Band Spectral Subtraction
I-SOS	Improved Spectral Over-Subtraction
ITU-T	International Telecommunication Union-Telecom sector
ISD	Itakura-Saito Distance
ISTFT	Inverse Short-Time Fourier Transform
ISS	Iterative Spectral Subtraction
LPF	Low pass Filter

LPC	Linear Predictive Coding
MBSS	Multi-Band Spectral Subtraction
MMSE	Minimum Mean Squared Error
MOS	Mean Opinion Score
MRA	Multi-Resolution Analysis
MMSE-STSA	Minimum Mean Square Error Short-Time Spectral Amplitude
NSS	Non-linear Spectral Subtraction
OLA	Overlaps-Add Method / Overlaps-Addition Method
PESQ	Perceptual Evaluation of Speech Quality
PAMS	Perceptual Analysis Measurement System
PSQM	Perceptual Speech Quality Measure
PM-SWPFB	Perceptually Motivated Stationary Wavelet Packet Filterbank
PM-SWPD	Perceptually Motivated Stationary Wavelet Packet Decomposition
PMS-MBSS	Perceptually Motivated Stationary Wavelet Packet Filterbank Based Multi-Band Spectral Over-Subtraction
QMF	Quadrature Mirror Filter
SNR	Signal-to-Noise Ratio
STSM	Short-Time Spectral Magnitude
STFT	Short-Time Fourier Transform
SOS	Spectral Over-Subtraction
SSF	Spectral Subtraction Filter
SSPP	Spectral Subtraction based on Perceptual Properties
SegSNR	Segmental Signal-to-Noise Ratio
SWPT	Stationary Wavelet Packet Transform
SWPFB	Stationary Wavelet Packet Filterbank
SWPD	Stationary Wavelet Packet Decomposition
VAD	Voice Activity Detector
WGN	White Gaussian Noise
WF	Wiener Filtering
WPT	Wavelet Packet Transform

Chapter 1

Introduction

1.1. Background

Speech is the most prominent and primary mode of interaction between human-to-human and human-to-machine communications in various fields such as automatic speech recognition and speaker identification [1]. The present day speech communication systems are severely degraded due to various types of interfering signals which make the listening task difficult for a direct listener and cause inaccurate transfer of information [2]. Therefore, to obtain near-transparent speech communication in applications such as in mobile phones, enhancement of degraded speech or equivalently the noise suppression has been one of the main research endeavors in the field of speech signal processing over the last few decades. The main focus of research in speech enhancement and its goals are :

i) to minimize the degree of distortion of the desired speech signal and to improve one or more perceptual aspects of speech, such as, the speech quality and/or intelligibility of the processed speech, i.e., to make it sound better or clearer to a human listener, which, in turn, results in the reduction of listening fatigue [3, 4];

ii) to improve the robustness of speech coders which tend to be severely affected by the presence of noise [5, 6] ; and

iii) to increase the accuracy of speech recognition system operating in less than ideal conditions [7, 8].

The first aim is by far the most common focus of most researches in the area. The quality of speech signal refers to how the signal is perceived by the listener, i.e., its pleasantness or comfort of listening. It may also include attributes such as naturalness which is of highly subjective nature [9, 10]. On the other hand, intelligibility, refers to what the speaker has said, in terms of meaning or information content or how much information can be extracted from a speech signal, is an objective measure [3, 9, 10]. Quality can be measured using the mean opinion score (MOS) where a listener rates the quality of the speech on five-point scale (i.e. from 1 to 5) [9, 10]. Intelligibility measurements are conducted differently as the emphasis is on the understanding of speech. In such listening tests, listeners are asked to listen to various sentences or isolated words, and they have to write the words they can recognize. Based on the percentage of correctly recognized words, an intelligibility score is obtained [11]. These two features, quality and intelligibility are however uncorrelated and independent of each other in a certain context. For example, a very clean speech of a speaker in a foreign language may be of high quality to a listener but at the same time it will be of zero intelligibility. Therefore, a high quality speech may be low in intelligibility while a low quality speech may be high in intelligibility [10]. An example of latter was shown in an earlier work by Thomas *et* at. [12], in which high pass filtering and clipping were applied to a noisy speech utterance. The high pass nature of the filter boosted the unvoiced portions (see Section 1.2.1) of the speech, which are crucial to the understanding of the words, thus improving the intelligibility of the filtered speech. However, the enhanced speech sounds distort significantly and the quality of the speech is degraded considerably. Therefore, it is very difficult to optimize both attributes together. In many earlier works, speech enhancement algorithms tend to increase the quality but used to reduce the intelligibility also. Newer algorithms function better and have less of such trade-off.

The second aim is important, as speech vocoders (which encode only the perceptual important aspects of speech with fewer bits) are highly affected by the presence of noise. This is because compressed speech is represented by a minimum number of bits and the previous few bits may be used for representing noise instead of speech. The presence of noise also affects the accurate analysis of the speech, resulting in faulty parameters, e.g. linear predictive coding coefficients. The effect of noise on coded speech is often many times worse than the original uncompressed noisy speech. Speech enhancement has been shown to be an important front-end processing for such systems [5, 6].

The third goal of speech enhancement is to address the well-known problem, increasing the accuracy of speech or a speaker recognition system in adverse conditions. It is well-known that the performance of such system degrades rapidly in practical noisy conditions [11]. This is due to the acoustic mismatch between the features used to train and test the systems and the ability of the models to describe the degraded speech. Typically, clean speech is used to train the acoustic models. Therefore, enhancement techniques which remove noise leaving an estimate of the clean signal are useful as a front-end (where speech enhancement is used as a pre-processing step for reducing the noise content and then use the enhanced signal as input to the speech processing system) [8, 24]. Put differently, speech enhancement can be regarded as an estimation problem, in which the clean signal is estimated from a set of noisy observations that consists of a clean speech signal plus random noise interferences [3, 13].

The noise sources present in a real environment can be of various types (see Section 1.2.1) and there is no general solution for all types of environments and applications. Further, there are many practical situations where speech processing systems are confronted with an adverse environment and therefore need a powerful enhancement algorithm. The aim of each particular application has to be taken into account while developing the suitable speech enhancement algorithm for the particular case. The thesis contains the study and development of various related speech enhancement algorithms which may find applications in adverse noisy conditions. Some of the possible applications are:

i) Digital mobile radio telephony: radio channel with variable transfer function, background noise (moving car) and limited bandwidth.

ii) Hands-free telephone systems.

iii) Telephones located in adverse environments such as factories, city streets, and aircraft cockpits etc.

iv) Teleconferencing systems.

v) Communications over noisy channel (long-distance telephone lines).

vi) Communication terminals in office environments (babble noise, computer fans, etc.).

vii) Hearing aids design: needs an effective noise reduction front-end along with speech processing module to suit the listener's disability [7].

Due to the long history of speech enhancement, various algorithms have been proposed. A broad classification of the approaches will be divided into temporal and transform based processing methods. In temporal processing, direct operations are done on the speech waveform. Filtering is performed directly on the time sequence which includes techniques such as linear predictive coding [14], Kalman filtering [15-17] etc. In the transform based processing, the noise reduction is performed on the transform coefficients [13, 18-21]. The various types of transforms are used such as Fourier transform, short-time Fourier transform [13, 18-24, 89-91], and wavelet transform [27, 30-36, 88]. The idea behind performing the transform is that it should be easier to distinguish between speech and noise in the transform domain. The transform based techniques seem to be more popular among many researchers.

Among various transform based speech enhancement algorithms, prsented in Section 1.2.3), the spectral subtraction method proposed by S.F. Boll [13], is one of the most widely used methods based on the direct estimation of short-time spectral magnitude (STSM). The main attraction of spectral subtraction method is:

i) Its relative simplicity, in that it only requires an estimate of the noise power spectrum, and

ii) Its high flexibility against the variation of subtraction parameters.

Despite its capability of reducing the background noise, spectral subtraction method [13] introduces perceptual noticeable spectral artifacts, known as remnant musical noise, which is composed of un-natural artifacts with random frequencies and perceptually annoys to the human ear. This noise is caused due to the inaccuracies in the short-time noise spectrum estimate and it faces difficulties in pause detection. In recent years, a number of speech enhancement algorithms have been developed to deal with the modifications of the spectral subtraction method to combat the problem of remnant musical noise artifacts and improve the quality of speech in noisy environments. These frequency domain speech enhancement algorithms constitute a family of spectral subtractive-type algorithms (see Section 2.5.1) and are based on subtracting the estimated short-time spectral magnitude of the noise from the STSM of noisy speech or by multiplying the noisy spectrum with gain functions and to combine it with the phase of the noisy speech [13, 20-24, 89-91].

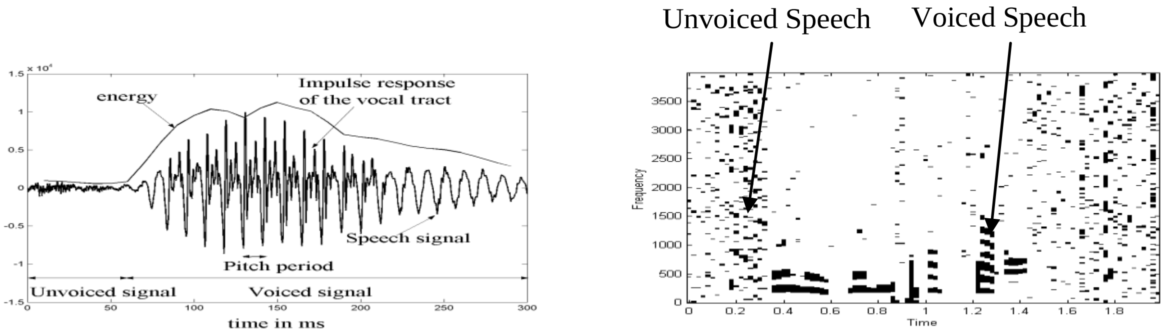
The wavelet transforms have been applied to various areas of research including speech and image de-noising, compression, detection, and pattern recognition, which can easily be computed by filtering a signal with multi-resolution filterbanks [25, 26]. In [27, 31], wavelet transform has been applied for enhancement of speech on the basis of a simple and powerful de-noising technique known as wavelet thresholding (shrinking). However, it is not possible to separate the speech signal from noises by a simple threshold because applying a uniform threshold to all wavelet coefficients would remove some speech components, as well, while suppressing additional noise. This is especially true for the non-stationary or colored noise degraded signal and some deteriorated speech conditions. The unvoiced components of speech are often eliminated from this method which results in loss of much a-tonic information that affects the quality of reconstructed signal [27]. Yet, it remains unclear which speech enhancement algorithm performs so well where the background noise level and characteristics are constantly changing.

In real-world environment, most of the noise captured by a microphone (see Section 1.2.3) is non-stationary or colored. If degrading noise is non-stationary, the multi-band (subband) approach has been found to be more efficient than whole band approach [3, 21]. In this thesis, we have presented transform based multi-band speech enhancement algorithms, especially suited for non-stationary conditions. In the next section, we describe the speech enhancement algorithms for adverse conditions.

1.2. Speech Enhancement in Adverse Environment

1.2.1 Description of Speech Signal

Speech is a non-stationary signal due to the time varying nature of the human speech production system. The information it carries is contained in variations of the signal in time and frequency. Furthermore, successive samples of the speech signal are highly correlated. In general, speech processing algorithms operate on a frame-by-frame basis with frame sizes ranging from 10 to 30 ms during which the signal is considered as stationary/quasi-stationary (short-time processing). A detailed description of speech properties can be found in [11, 28], nevertheless, few characteristics can be outlined here. Speech bandwidth varies approximately from 300 Hz to 3.4 kHz, but some consonants such as, 'f', 's', 't', have frequency component up to 8 kHz. Speech can be decomposed in two principal classes of sounds: voiced speech, which has higher amplitude and energy at low frequencies (below 3 kHz) and unvoiced speech, with lower energy but spreading to higher frequencies. Voiced speech is produced when periodic pulses of air generated by the vibrating glottis resonate through the vocal tract, at frequencies dependent on the vocal tract shape. Unvoiced speech is non-periodic, random-like sounds, caused by air passing through a narrow constriction of the vocal tract as when consonants are spoken. There are also a small number of speech sounds which employ mixed excitation, such as voiced fricatives 'z' and 'v', as well as voiced stops as 'd' and 'b'. In a speech signal about two-thirds of speech is voiced and one-third of speech is unvoiced. Therefore, voiced speech seems to be more important than unvoiced speech for preserving speech quality. Although unvoiced sections are shorter and of lower energy but they are important for speech perception as well and most important for intelligibility. An example of unvoiced and voiced speech is given in Fig. 1.1.



(a)

(b)

Fig. 1.1: Examples of unvoiced and voiced speech: (a) time domain, and (b) time-frequency domain (spectrogram).

1.2.2 Noise Characteristics

Noise can be defined as an unwanted signal and there are many forms of noise. The choice of particular speech enhancement method depends on the nature of the noise. Therefore, a good model of the noise source is important to analyze how well a speech enhancement algorithm/model works with different types of noise. Noises can have different statistical, spectral or spatial properties, as summarized in Table 1.1.

TABLE 1.1. CLASSIFICATION OF NOISE BASED ON VARIOUS PROPERTIES.

Properties	**Types of Noise**
Structure	Continuous/Impulsive/Periodic
Types of interaction	Additive/Multiplicative/Convolutive
Temporal behaviour	Stationary/Non-stationary
Frequency range	Broadband/Narrowband
Signal Dependency	Correlated/Un-correlated
Statistical properties	Dependent/Independent
Spatial Properties	Coherent/Incoherent

Depending on the nature and the properties of the noise source, we can define the following classification [11, 29]:

1) Background noise: additive, uncorrelated noise present in a lot of environments (apart from a sound proof room), such as cars (engine, road noise), offices, fans, city streets, machines, traffic noise, factories, aircraft cockpit. For example, operating a hands-free mobile phone in a car can be affected by at least three different types of background noises, namely wind, road as well as engine noise. It can be stationary, slowly varying or non-stationary. The speech signal degraded by additive uncorrelated background noise is defined as noisy (unprocessed) speech [29].
2) Interfering speaker (speech-like noise): additive noise is composed of one or more competitive talkers. If the noise is a multi-talker babble noise, the phenomenon is called the 'cocktail party effect'. This noise has the characteristics and frequency range similar to that of the useful signal.
3) Noise correlated with the signal : for example reverberations, echoes.
4) Impulse noise: for example, noise resulting from slamming doors, noise in old recordings.
5) Non-additive noise: transmission noise or channel distortion (speech on the telephone line), spectral shaping and non-linearities due to microphones.
6) Non-additive noise due to speaker stress: noise can also induce changes in the speech production process.

The most difficult situation to handle is a general non-stationary noise when no *a-priori* knowledge is available about the noise characteristics. Furthermore, if there is a temporal, frequency or spatial overlap between noise and speech, it is difficult to reduce the noise without distorting the speech. Some of the popular noise databases used for research in the area of speech are NOIZEUS[!], TIMIT[!!].

[!] NOIZEUS - Noisy Speech Corpus. The noise database was prepared from the AURORA database.

[!!] TIMIT Acoustic - Phonetic Continuous Speech Corpus. Corpus design was a joint effort between the Massachusetts Institute of Technology (MIT) and Texas Instruments (TI).

1.2.3 Classification of Speech Enhancement Methods

There are several methods proposed for speech enhancement in past decades. The reported algorithms can be categorized into two main classes as parametric and non-parametric methods. Parametric approaches needs a mathematical model for the signal generation or production process. This model describes the predictable structures and the observable patterns in the process. In this approach, noise suppression is performed depending on this *a-priori* information. Since the enhancement is based on the parametric model, selection of the model is crucially important in these algorithms. Whereas, non-parametric methods does not require the speech generation or production system. It simply requires an estimation of the noise spectrum. The noise spectrum can be estimated from the silence periods where the speaker is silent (single channel) or from a reference source (multi-channel) [29].

The classifications of speech enhancement methods also depend on the number of available microphones that are used for recording speech data, namely, single, dual and multi-channel approaches. In case of single channel, only one microphone input is available and therefore only one mixture (noisy) speech signal. This is the most difficult situation, since the noise and the speech are in the same channel. This situation is especially difficult in two cases: i) when speech and noise overlap in the time and frequency domain, and ii) when noise is non-stationary. The single channel is especially useful in mobile communication applications, where only a single microphone is available due to cost and size considerations. In case of multi-channel, several microphones are spatially distributed in the surrounding, leading to more mixture speech signal which exhibits the advantage of both spatial and spectral information. However, since multi-microphones will come at an increased cost and may not be always available, the single channel speech enhancement always attracts attention [9].

The basic assumptions of our work are the following: i) non-parametric and single channel speech enhancement, ii) additive and uncorrelated noise, and iii) low SNR (<10 dB). Fig. 1.2 shows the basic overview of additive noise and a speech enhancement system (speech enhancement model). Here $s(n)$, $y(n)$ and $\hat{s}(n)$ are the clean speech signal, noisy speech signal and the enhanced speech signal, respectively. The source of degradation is an additive random interference/noise $d(n)$ and the resulting degraded/noisy speech $y(n) = s(n) + d(n)$. Here, it is assumed that the noise is additive and statistically independent with the signal. A single channel system cannot capture the time variation of noise, it is often impossible to suppress noise without distorting speech. As a consequence, the performance of such systems is limited by a trade-off between speech distortion and noise reduction.

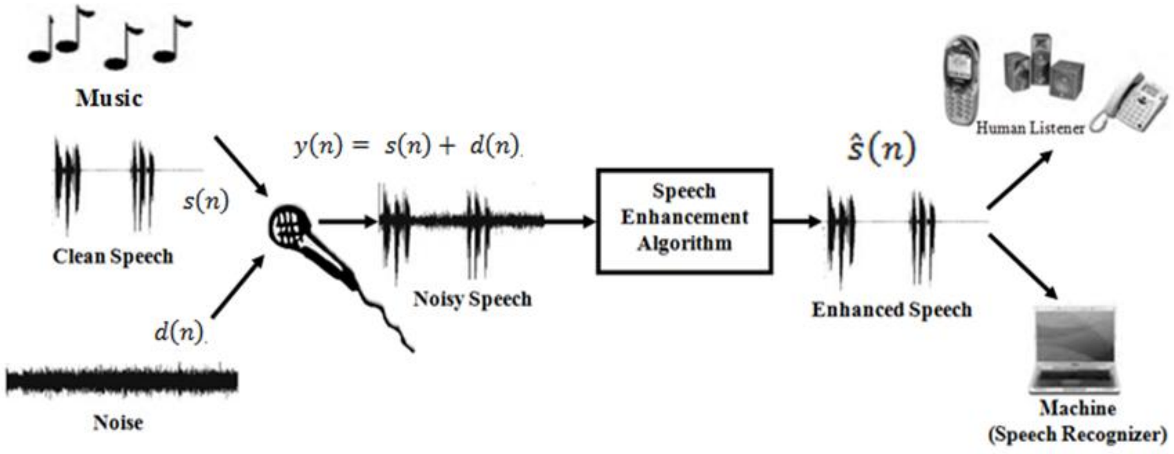


Fig. 1.2: Basic overview of additive noise and a speech enhancement system (speech enhancement model).

1.3. Motivation and Objective of the Thesis

In noisy environment, the background noise, as described in Section 1.2.2, degrades the quality and intelligibility of the perceived speech and thus decreases the efficiency of communication between humans due to reduced intelligibility and quality. Also, the performance of speech enhancement systems significantly affects various speech processing systems, such as speech coding and recognition where the speech enhancement is routinely done in the pre-processing stage. Therefore, suppression of background noise is a relevant problem.

The notable speech enhancement algorithms for suppressing background noise include spectral subtraction method [13, 20], Wiener filtering [22] and signal subspace based methods [3]. Among these algorithms classical spectral subtraction method has been widely use for enhancement of degraded speech. The popularity of the spectral subtraction method is because of its simplicity and computational efficiency [3, 4]. In spectral subtraction method, enhanced speech is obtained by estimating the noise spectrum from the noisy speech and by subtracting the estimate of noise from the noisy speech spectrum [3, 4].

This spectral subtraction method works well for additive stationary noise [3, 4]. The main problem of this method is the introduction of remnant musical noise which accompanies in the enhancement process. The remnant noise is a perceptually annoying noise with random frequencies and of irritating nature. The remnant noise can be reduced to some extent by introducing a set of subtraction parameters in the algorithm that also distorts the estimated speech [20]. Further, the drawback of this classical approach and its derivatives is that it uses a fixed set of values for the subtraction parameters. Additionally, these algorithms are limited to restrictive applications and generally work well with white Gaussian noise.

The real-world noise is mostly non-stationary in nature, where it affects the speech spectrum non-uniformly [3]. Thus, a multi-band variation of spectral subtraction is found to be more appropriate, [3, 21, 47-50]. In multi-band approach, the speech is decomposed in a set of bands and spectral subtraction is employed in each band separately. Although, the multi-band spectral subtraction (MBSS) algorithm gives a certain level of performance improvement, the presence of remnant noise still exists and affects human listening. To address this problem, we have proposed several improvements on the multi-band spectral subtraction algorithms and studied their impact on speech enhancement.

The first improvement on the multi-band spectral enhancement is the proposition of iterative processing of the enhanced speech. It is known that remnant noise the estimated speech can be suppressed further by iterating the base algorithm to a limited number of times for the whole spectrum [23, 51-54, 77]. Thus, if we iterative the enhanced speech obtained by MBSS to a finite number of time, further reduction in remnant noise can be obtained. Therefore, an iterative processing can be employed for the multi-band case, where the remnant noise is estimated from the enhanced speech from each band separately, and the operation is repeated to a finite number of times. This approach of multi-band iterative processing for speech degraded by non-stationary noise is likely give us further improvement in terms of speech quality. In this thesis, we have explored an iterative processing based multi-band spectral subtraction algorithm.

Next, we incorporate the perceptual frequency scale of human hearing for the processing of speech signals for developing the multi-band enhancement algorithms. This is because, the human listeners are the final judge to evaluate the quality of the enhanced speech and the selection of frequency bands in accordance with the perceptual auditory system is expected to give us better performance [24, 55]. The decomposition and processing of speech signal in resemblance with the human auditory system has been utilized in many speech processing applications and substantial performance improvement has been achieved [24, 30, 32-35, 55, 66, 67]. The perceptually motivated non-uniformly spaced frequency bands have been employed for the proposed speech enhancement algorithms in our work.

We have also explored the importance of precise estimation of noise spectrum for the accurate performance of these enhancement algorithms. Usually, the noise estimation is done by detection of speech pauses/silences using voice activity detector for identifying the segments of pure noise [3, 78-81]. In practical situation, this is a difficult task, especially if the background noise is non-stationary in nature or the signal-to-noise ratio (SNR) is low. Thus, a voice activity detector may not identify the noisy speech segments signal at very low SNR [3]. Hence, there is a strong need to update the noise spectrum adaptively and continuously over time [46, 82-84, 86]. In our work, an adaptive noise estimation approach has been employed for the estimation of noise that does not require the explicit detection of voice activity. The noise estimate is updated by adaptively smoothing the noisy signal power and the smoothing parameter is controlled by a linear function of estimated SNR.

The speech signal is non-stationary in nature and contains frequently transients. For analyzing non-stationary signals, such as speech, wavelet transform have been proved to be is an important tool and has been used in various speech applications [30-36]. Thus, the spectral subtraction algorithm with wavelet transform based time-frequency decomposition is likely to give better performance than methods employing short-time Fourier transform (STFT). Also, the extension of discrete wavelet transform, namely, wavelet packet transform (WPT), is suitable for matching the frequency bands closely with the auditory frequency scale [25, 26, 57-61]. The over-sampled filterbank realization of WPT i.e., stationary WPT (SWPT) [62-68] has been explored in this thesis as the front-end transform for multi-band speech enhancement algorithm.

In this thesis, we have proposed speech enhancement algorithms and explored the implications of the iterative processing, improved noised estimation technique and the non-uniform frequency decomposition in accordance of human auditory system in the framework of multi-band spectral subtraction approach. The time-frequency decomposition of the proposed algorithms is based on the STFT and SWPT. The iterative processing based multi-band spectral subtraction algorithm is proposed first. Next, we have developed the non-uniformly frequency spaced multi-band speech enhancement algorithms based on perceptual frequency scale of human auditory system. The non-uniformly spaced multi-band approach makes use of both the transforms, namely STFT and SWPT and employs the proposed adaptive noise estimation technique.

1.4. Major Contributions and Organization of the Thesis

In this thesis, we have developed and implemented three transform domain multi-band spectral subtraction based single channel speech enhancement algorithms which exploit the iterative processing, adaptive noise estimation technique and perceptual frequency scale of human auditory system. In the iterative processing based multi-band spectral subtraction algorithm spectral, the speech is processed into four uniformly spaced frequency bands and spectral subtraction is performed independently on each band using band-specific over-subtraction factor This iteration process is iterated a small number of times and noise is estimated in each iteration. The next speech enhancement algorithm is an improved multi-band spectral subtraction based on critical band scale of human ear. Here, the noisy speech is processed by splitting the spectrum into six non-uniformly spaced frequency bands in accordance to the critical band rate scale and the spectral subtraction is performed independently on each band using band-specific over-subtraction factor. The proposed algorithm utilizes an adaptive noise estimation technique for estimating the noise in each band which does not need explicit speech silence detection. The third speech enhancement algorithm utilizes a wavelet transform based approach to decompose the degraded speech signal. The proposed algorithm **is** a perceptually motivated stationary wavelet packet filterbank (PM-SWPFB) based improved multi-band spectral over-subtraction (I-SOS) algorithm for the enhancement of narrowband speech degraded by non-stationary noises. The research work contained in this thesis is summarized as follows:

- Determination of the limitations of the single channel speech enhancement techniques.
- Extensive study and comparison of the spectral subtractive-type speech enhancement algorithms. Unification of the formulation.
- Analysis and comparison of the structure of the remnant musical noise for various speech enhancement algorithms.
- Study of wavelet transform, wavelet packet transform and stationary wavelet packet transform and their filterbank implementations.
- Proposition of an iterative processing based multiband speech enhancement algorithm.
- Realization of a noise estimation algorithm in an adverse environment.
- Proposition of the transform based (STFT and SWPT) multi-band speech enhancement algorithms driven by the auditory perception criterion.
- Study of the spectrogram and traditional objective measures along with the perceptually motivated objective measures for the analysis of the performance of the proposed enhancement algorithms and their correlation with subjective listening tests using the NOIZEUS speech corpus.

This thesis is divided into seven chapters, including this introduction chapter and the rest of the thesis is organized as follows:

Chapter 2 presents an extensive study of the family of enhancement algorithms, known as, spectral subtractive-type algorithms. This study presents a comparison of various existing spectral subtraction algorithms and their derivatives. The methods used to reduce the remnant noise, which is a major drawback of these algorithms are presented with the results. This chapter forms the basis for the enhancement algorithms developed in the later chapters of this thesis. A study of the performance limitation of single channel enhancement algorithms is also carried out. **Chapter 3** presents the complete overview of stationary wavelet packet transform which is utilized for the speech enhancement algorithm developed in Chapter 6. It starts with a brief review of the properties of wavelet, wavelet packet transforms, and stationary wavelet packet transforms. This chapter also illustrates how to use the filterbank structures to implement wavelet transform based speech signal processing. The content of this chapter forms the basis for the development of perceptually motivated stationary wavelet packet filterbank based speech enhancement algorithm developed in Chapter 6 of this thesis.

The major contributions of this thesis are presented in **Chapter 4, Chapter 5**, and **Chapter 6,** each of which proposes an enhancement algorithm and evaluates their performance with several objective measures as well as subjective measures separately.

In **Chapter 4,** a multi-band spectral subtraction algorithm based on iterative processing is proposed for the enhancement of degraded speech. This algorithm aims to give us additional remnant noise suppression over and above the background noise reduction obtained from the traditional multi-band approaches. In this technique, the speech is processed into non-overlapped uniformly spaced frequency bands, numbering four and spectral subtraction is performed independently in each band using band-specific over-subtraction factor. This process is iterated a small number of times and the noise is estimated for each band in each iteration. The central idea of this algorithm is that after the completion of the typical enhancement process the additive noise transforms into the remnant noise and afterwards, the remnant noise can be estimated in each iteration.

In **Chapter 5,** an improved multi-band spectral subtraction algorithm based on critical band rate scale is proposed for the enhancement of speech degraded by non-stationary noises. Here, the narrowband speech is processed by splitting into six non-uniformly spaced frequency bands in accordance to the critical band rate scale. Subsequently, the spectral subtraction is performed separately in each band using over-subtraction factors computed for the bands. The proposed algorithm uses an adaptive noise estimation technique to estimate the noise power in each band without the requirement of explicit speech pause detection.

Chapter 6 proposes a perceptually motivated stationary wavelet packet transform based improved multi-band spectral over-subtraction algorithm for the enhancement of narrowband speech degraded by non-stationary noises. The perceptually motivated stationary wavelet packet filterbak is obtained from the uniformly spaced stationary wavelet packet tree that closely mimics the critical bands of the perceptual auditory system. After the decomposition of the input noisy speech signal into seventeen non-uniform subbands by the filterbank, the improved spectral over-subtraction algorithm is used to estimate the speech from each subband. The spectral over-subtraction approach over uses the noise estimation technique similar to the one as proposed in Chapter 5, to estimate noise power from each subband without the need of separating the speech from non-speech regions. The noise estimate in each subband is updated by adaptively smoothing the noisy signal power.

We have used the NOIZEUS speech corpus for performance evaluation of the proposed algorithms in Chapter 4, 5 and 6 in terms of the quality of the enhanced speech. The spectrogram analysis is done for these algorithms and various objective measures such as signal-to-noise ratio (SNR), segmental SNR (SegSNR), Itakura-Saito distortion (ISD) and perceptual evaluation speech quality (PESQ) are obtained on the test data. Several real-world noises, such as, car noise, train noise, restaurant noise, babble noise, airport noise, street noise, exhibition noise and a computer generated stationary noise at various different levels of SNRs, have been taken for the evaluation purpose. A subjective listening test based on the mean opinion score (MOS) is also carried out and the results are compared with be objective measure of subjective speech quality, PESQ.

Finally, **Chapter 7** concludes this thesis with a summary of the main developments and the contributions. Further, the chapter contains a discussion about the possible direction of future research work related to the work contained in the thesis.

Chapter 2

Spectral Subtractive-Type Algorithms

2.1. Introduction

The spectral subtraction method is a classical approach for enhancement of single channel speech degraded by additive background noise. The basic principle of this method is to estimate the short-time spectral magnitude of speech by subtracting estimated noise spectrum from the noisy speech spectrum [13, 19, 38, 89-91]. This is achieved by multiplying the noisy spectrum with a gain function and later combining it with the phase of the noisy speech [38, 39]. The main drawback of this method is the presence of distortions in the enhanced speech, which is caused due to random variations of noise having a musical structure, called remnant musical noise [20]. Many derivatives of this method have been developed over the past years to address these limitations [20-24]. These variations constitute a family of spectral subtractive-type algorithms. The aim of this chapter is to provide a comparative and simulation study of the different forms of spectral subtractive-type algorithms. The study conducted in this chapter forms the basis of the algorithms proposed in this thesis in Chapter 4, Chapter 5 and in Chapter 6.

The work reported in this chapter has resulted in the following publications

[I]. Navneet Upadhyay, and Abhijit Karmakar, "Spectral subtractive-type algorithms for enhancement of noisy speech: an integrative review," *International Journal of Image, Graphics and Signal Processing,* 2013. [Accepted]

[II]. Navneet Upadhyay, and Abhijit Karmakar, "The spectral subtractive-type algorithms for enhancing speech in noisy environments," in *Proceedings of Ist IEEE International Conference on Recent Advances in Information Technology*, ISM Dhanbad, India, March 15─17, 2012, pp. 841─847.

[III]. Navneet Upadhyay, and Abhijit Karmakar, "A frequency dependent algorithm for enhancement of degraded speech," presented in National Conference on VLSI Design and Embedded Systems, CEERI Pilani, India, Oct. 12─14, 2011.

The algorithms that are taken for the comparative study include the basic magnitude spectral subtraction (BSS) algorithm developed by Boll [13], spectral over-subtraction (SOS) algorithm [20], multi-band spectral subtraction (MBSS) algorithm [21], Wiener filtering (WF) [22], iterative spectral subtraction (ISS) [23], and spectral subtraction based on perceptual properties (SSPP) [24]. It is evident from the simulations and evaluations results that the modified forms of spectral subtraction method reduces remnant musical noise efficiently, and the enhanced speech, thus obtained, has minimal speech distortion with improved signal-to-noise ratio.

The rest of Chapter 2 is structured as follows. In Section 2.2, the principle of spectral subtraction method is described. In Section 2.3, the limitation of spectral subtraction method is explained followed with two sub-sections, Section 2.3.1 and Section 2.3.2. Section 2.3.1 explains the remnant musical noise and Section 2.3.2 describes the speech distortion. In Section 2.4, noise estimation techniques are discussed. In Section 2.5, an improvement to spectral subtraction method is discussed followed by sub-section Section 2.5.1, which presents a detailed study of spectral subtractive-type algorithms. Section 2.6 presents the implementation, experiments results, and performance evaluation of spectral subtractive-type algorithms. Finally, Section 2.7 concludes this Chapter.

2.2. Principle of Spectral Subtraction Method

The spectral subtraction method is one of the most popular and computationally simple methods for effectively suppressing the background noise from the noisy speech as it involves a single forward and inverse transform. The first comprehensive spectral subtraction method, proposed by Boll [13] is based on non-parametric approach, which simply needs an estimate of noise spectrum and used for both speech enhancement and speech recognition.

In real-world listening environments, the speech signal is mostly degraded by additive noise [2, 3, 13, 29]. Additive noise is typically the background noise and is uncorrelated with the clean speech signal. The background noise can be of stationary, such as white Gaussian noise (WGN) or of non-stationary or colored, such as multi-talker (babble) noise, restaurant noise, car noise etc. People talking in the background also contribute to the unwanted noise in the degraded speech signal. The signal degraded by background noise is termed as noisy speech. The noisy signal can be modeled as the sum of the clean speech signal and the random noise [2, 3, 13, 29, 90, 91] as

$$y(n) = s(n) + d(n), \; 0 \leq n \leq N-1 \tag{2.1}$$

where n is the discrete time index and N is the number of samples in the signal. Here, $y(n)$, $s(n)$, and $d(n)$ are the n^{th} sample of the discrete time signal of noisy speech, clean speech and the noise, respectively. Since, the speech signal is non-stationary in nature and contains transient components, usually the speech signal is divided in small frames to make it stationary or quasi-stationary over the frames and short-time Fourier transform (STFT) is used for further processing. The STFT is of fundamental importance to signal analysis as it introduces a time dependent frequency analysis, which is not provided by the Fourier transform. The STFT is obtained by applying the Fourier transform at different points in time on finite length (i.e., frame) sections of a signal. Now representing the STFT of the time windowed signals by $Y_W(\omega)$, $D_W(\omega)$, and $S_W(\omega)$, (2.1) can be written as [3, 13, 29],

$$Y_W(\omega) = S_W(\omega) + D_W(\omega) \tag{2.2}$$

where ω is the discrete frequency index of the frame and W is the window (Hamming or Henning window). Put differently, for implementation of spectral subtractive-type speech enhancement algorithms, few assumptions are necessary. First, the speech signal should be stationary; second, the noise is zero mean and uncorrelated with clean speech signal [37]. Throughout this thesis, it is assumed that the signal is segmented into frames first and then windowed, hence for simplicity, we drop the use of subscript W from windowed signals.

The spectral subtraction method mainly involves two stages. In the first stage, an average estimate of the noise spectrum is subtracted from the instantaneous spectrum of the noisy speech. This is termed as basic spectral subtraction step. In the second stage, several modifications like half-wave rectification (HWR), remnant noise reduction and signal attenuation are done to reduce the signal level in the non-speech regions. In the entire process, the phase of noisy speech is kept unchanged because it is assumed that the phase distortion is not perceived by human auditory system (HAS) [38, 39]. Therefore, the short-time spectral magnitude (STSM) of noisy speech is equal to the sum of STSM of clean speech and STSM of random noise without the information of phase and (2.2) can be expressed as

$$|Y(\omega)| = |S(\omega)| + |D(\omega)| \tag{2.3}$$

where $Y(\omega) = |Y(\omega)| \exp(j\varphi_y(\omega))$, $S(\omega) = |S(\omega)| \exp(j\varphi_y(\omega))$, $D(\omega) = |D(\omega)| \exp(j\varphi_y(\omega))$ and $\varphi_y(\omega)$ is the phase of the noisy speech. To obtain the short-time spectrum of noisy speech $Y(\omega)$ is multiplied by its complex conjugate $Y^*(\omega)$. In doing so, (2.2) become

$$|Y(\omega)|^2 = |S(\omega)|^2 + |D(\omega)|^2 + S(\omega)D^*(\omega) + S^*(\omega)D(\omega) \tag{2.4}$$

Here $D^*(\omega)$ and $S^*(\omega)$ are the complex conjugates of $D(\omega)$ and $S(\omega)$, respectively. The $|Y(\omega)|^2$, $|S(\omega)|^2$, and $|D(\omega)|^2$, are referred to as the short-time spectrum of noisy speech, clean speech, and random noise, respectively. In (2.4), the value of $|D(\omega)|^2$, $S(\omega)D^*(\omega)$ and $S^*(\omega)D(\omega)$ cannot be obtained directly and are approximated as, $\mathrm{E}\{|D(\omega)|^2\}$, $\mathrm{E}\{S(\omega)D^*(\omega)\}$ and $\mathrm{E}\{S^*(\omega)D(\omega)\}$, where $\mathrm{E}\{.\}$ denotes the ensemble averaging operator. As the additive noise is assumed to be zero mean and uncorrelated with the clean speech signal, the terms $\mathrm{E}\{S(\omega)D^*(\omega)\}$ and $\mathrm{E}\{S^*(\omega)D(\omega)\}$ reduce to zero. Therefore, (2.4) can be rewritten as

$$\begin{aligned} |\hat{S}(\omega)|^2 &= |Y(\omega)|^2 - \mathrm{E}\{|D(\omega)|^2\} \\ &= |Y(\omega)|^2 - |\widehat{D}(\omega)|^2 \end{aligned} \tag{2.5}$$

where $|\hat{S}(\omega)|^2$ and $|Y(\omega)|^2$ is the power spectrum of estimated speech and the noisy speech, respectively. Here term $|\widehat{D}(\omega)|^2$ is the average noise power, normally estimated and updated during speech pauses using voice activity detector (VAD), as explained in Section 2.4. A VAD is used to discriminate between voice activities (i.e. speech presence) and silence/pause (i.e. speech absence). This assumption is valid for the case of stationary noise in which the noise spectrum does not vary over time.

In spectral subtraction method, it is assumed that the speech signal is degraded by additive white Gaussian noise (AWGN) with flat spectrum; hence the noise affects the signal uniformly over the spectrum. In this method, the subtraction process needs to be carried out carefully to avoid any speech distortion. The spectra obtained after subtraction process may contain some negative values due to inaccurate estimation of the noise spectrum. Since the power spectrum of estimated speech can become negative due to over-estimation of noise, but to get rid of this possibility, therefore, a half-wave rectification (by setting the negative portions to zero) or full-wave rectification (absolute value) are introduced. But the half-wave rectification (HWR) introduces annoying noise in the enhanced speech. Whereas, full-wave rectification (FWR) avoids the creation of annoying noise, but it is less effective in suppressing noise. Therefore, HWR is often used in spectral subtraction method due to its superior noise suppression ability. Thus, the complete power spectral subtraction algorithm is given by

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - |\widehat{D}(\omega)|^2 , & \text{if } |Y(\omega)|^2 > |\widehat{D}(\omega)|^2 \\ 0 & \text{else} \end{cases} \tag{2.6}$$

As the human perception is relatively insensitive to phase [38, 39], the enhanced speech spectrum can be obtained with phase of noisy speech and thus the reconstruction is done by taking the inverse STFT (ISTFT) of the enhanced spectrum using the phase of the noisy speech and overlap-add (OLA) method [19, 40] (see Appendix A), and can be expressed as

$$\hat{s}(n) = \text{ISTFT}\,\{|\hat{S}(\omega)|\exp(j\varphi_y(\omega))\} \tag{2.7}$$

On the contrary, a generalized form of spectral subtraction method (2.5) can be obtained by altering the power exponent from 2 to b, which determines the sharpness of the transition.

$$|\hat{S}(\omega)|^b = |Y(\omega)|^b - |\widehat{D}(\omega)|^b, b > 0 \tag{2.8}$$

where $b = 2$ represents the power spectrum subtraction and $b = 1$ represents the magnitude spectrum subtraction. In Fig. 2.1, the block diagram of basic spectral subtraction method is shown.

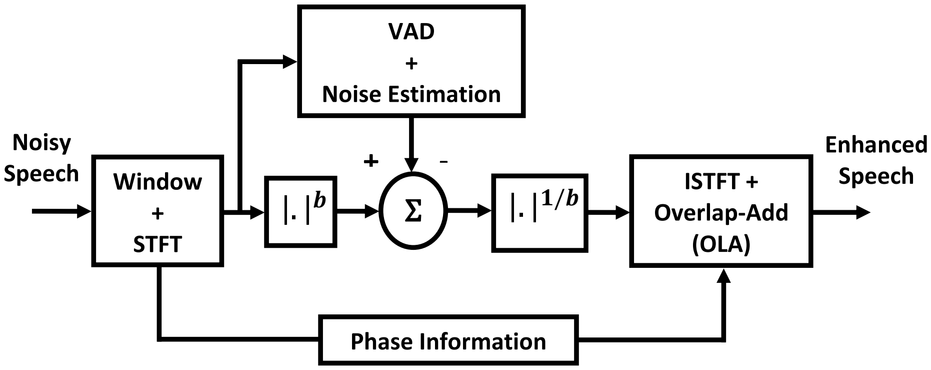


Fig. 2.1: Block diagram of spectral subtraction method.

2.3. Limitations of Spectral Subtraction Method

Although the spectral subtraction method is computationally simple and efficient for stationary or slowly varying broadband additive noise but it suffers from some severe drawbacks. From (2.5), it is clear that the effectiveness of spectral subtraction is heavily dependent on accurate noise estimation, which additionally is limited by the performance of speech/pause detectors [13]. A VAD performance degrades significantly at lower SNR. When the noise estimate is less than perfect, two major problems occur, such as, remnant noise, referred as musical noise, and speech distortion. These are discussed in the following sections.

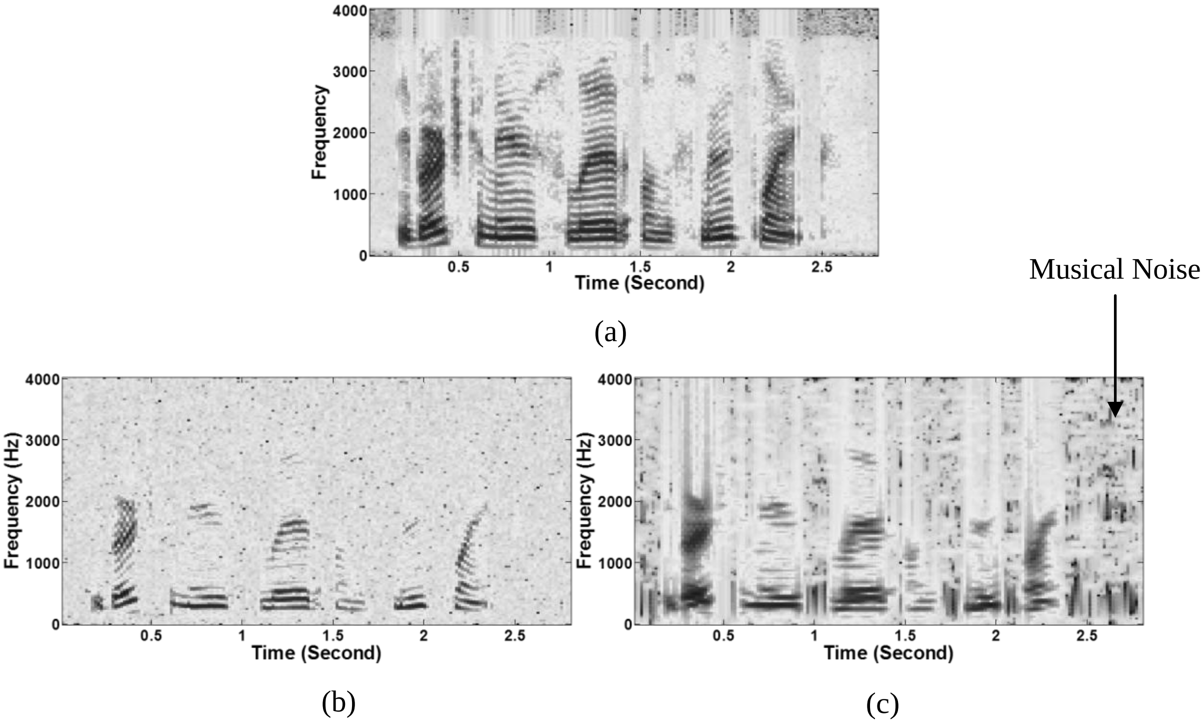
2.3.1 Remnant Musical Noise

When the average noise is subtracted from the noisy frames and the half-wave rectification is done, the remnant noise will have a magnitude between zero and a maximum value measured during speech pauses. Transformed back to the time domain, the noise will sound like sum of tone generators with random frequencies. Put differently, remnant noise is the noise remaining after the enhancement process, given by

$$r(n) = \hat{s}(n) - s(n) \qquad (2.9)$$

where, $r(n)$ is a remnant noise that has a musical nature leading to an unnatural quality. This effect is perceptually annoying because the remnant noise has a large variance and a 'musical' structure. A detailed study of this phenomenon is presented in [20]. Remnant noise is due to the subtraction of an averaged noise spectrum from an instantaneous spectrum, with values above and below the mean. It is of lower energy than the original noise, but with a higher variance due to tones at random frequencies. Therefore, this noise is very different from the original noise and can sometimes be even more disturbing. Remnant noise has an effect not only on human listener, but also on speech processing system like speech recognizer or speech coders. Since, the ultimate goal of speech enhancement is to provide good quality speech for human listeners, this is a severe flaw.

In Fig. 2.2, the spectrogram of clean speech, noisy speech, and enhanced speech are shown. On comparing the parts in Fig. 2.2 (a), (b) and (c), it is evident that the enhanced signal (c) contains random frequencies that are not present in the clean signal (a). This is an example of remnant musical noise.



(a)

(b)

(c)

Fig. 2.2: Speech spectrogram: (a) clean speech, (b) noisy speech, and (c) enhanced speech, respectively.

2.3.2 Speech Distortion

The second problem created by the speech enhancement algorithms is the speech distortion, also examined in [13]. This happens due to imperfections in the noise estimation process, when speech components are incorrectly attenuated or completely removed. This reduces the naturalness and intelligibility of the speech, and is again annoying to the listener.

As the spectral subtraction method gives reasonable quality speech with a good level of noise reduction, it has been the goal of speech enhancement researchers over the past two decades to find accurate noise estimation techniques and to minimize these errors [13]. Since, it is not possible to completely eliminate both; researchers have tried to reduce distortion, while keeping the remnant noise below a certain level, as an acceptable trade-off between the two. It has been proven that listeners can tolerate a low level of noise, provided that it is of the same spectrum as the original additive noise, as this is less irritating than random musical noise [24].

2.4. Noise Estimation

Most of the single channel enhancement systems need an estimation of the noise spectrum. Noise estimation is usually performed during speech silences/pause (see Section 2.4.1) segments of the speech signal. However, the speech/silence detection is not always reliable at low SNRs. This assumption is valid for the case of stationary/quasi-stationary noise, where the noise spectrum does not vary over time. Traditional VADs track the noise only frames of the noisy speech to update the noise estimate. But the update of noise estimate in those methods is limited to speech silence frames. Furthermore, if the noise is non-stationary in which the power spectrum of noise varies even during speech activity, it is not sufficient to update the noise estimate during speech silence, and therefore the system is unable to track the non-stationarities of noise. To overcome this effect, methods that are able to perform noise estimation during speech activity have been proposed. They are described in Section 2.4.2.

2.4.1 Estimation during Speech Silences

If noise is estimated during non-speech periods, these periods have to be long enough to obtain a good estimate with a small variance. Furthermore, this kind of noise is conditioned by the existence of a robust speech/noise detector.

The noise estimate can be obtained by averaging squared spectral amplitudes during speech silences. Generally, it is updated on a frame-by-frame basis according to the exponential averaging calculated at frame k [29, 41]:

$$|\widehat{D}(\omega,k)|^2 = \lambda_D.|\widehat{D}(\omega,k-1)|^2 + (1-\lambda_D).|Y(\omega,k)|^2, 0.5 \leq \lambda_D \leq 0.9 \qquad (2.10)$$

where λ_D is a time and frequency dependent smoothing parameter whose value depends on the noise changing rate and k refers to the current frame index. Here, $|Y(\omega,k)|^2$ represents the short-time power spectrum of noisy speech, $|\widehat{D}(\omega,k)|^2$ is the updated noise spectral estimate, and $|\widehat{D}(\omega,k-1)|^2$ is the past noise spectral estimate. This equation is updated only during speech silence segments when $Y(\omega,k) = D(\omega,k)$. The approximate value of λ_D has to be chosen depending on the stationarity of the noise. This choice determines the number of noise frames τ used for the averaging. A typical value of λ_D for 20 ms frame is 0.9 that results in a time constant of about 10 frames, or 200 ms. The relation between τ and λ_D is approximately given in as [42]

$$\tau = \frac{2}{(1-\lambda_D)} \qquad (2.11)$$

2.4.2 Estimation during Speech Activity

Some methods have been developed for calculating the noise power estimate during speech activity, in order to track the non-stationarities of noise. Noise estimation methods that avoid explicit silence detection can be based on several principles:

i) Estimation of noise during unvoiced periods.

ii) Estimation of noise based on histograms.

iii) Estimation of noise by detection of valleys in the spectrum.

iv) Exploitation of the short-time characteristics of the speech signal.

A very simple method is presented in [43]. The noise estimate, calculated using (2.10), is updated in each frame (instead of updating the estimation only during speech pauses). However, the new estimate $|\widehat{D}(\omega,k)|$ has to be lower than 1.006 $|\widehat{D}(\omega,k-1)|$ and greater than 0.978 $|\widehat{D}(\omega,k-1)|$. This means that the estimate cannot increase faster than 3 dB or decrease faster than 12 dB per second. Therefore, the noise estimate will slightly increase during speech segments but will rapidly return to the correct value during noise segments.

Two methods avoiding explicit silence detection are described in [44]. The first method has a very low computational complexity and is also based on (2.10). When the value of $|Y(\omega,k)|$ is greater than the adaptive threshold $\mathrm{Th}(\omega)$ given by:

$$\mathrm{Th}(\omega)=\beta.|\widehat{D}(\omega,k-1)|, 1.5<\beta<2.5 \tag{2.12}$$

then the segment is considered as speech and it is not taken into account in the averaging described in (2.10).

The second method is based on the same principle, except that when $Y(\omega,k)$ is below the threshold $\mathrm{Th}(\omega)$, the noise level is estimated from histograms of the past values (about 400 ms) of $|Y(\omega,k)|$ in 40 frequency bins, instead of using (2.10). The noise estimation is obtained by calculating the maximum in each band, with a smoothing versus time. Evaluation of histograms leads to more accurate results than exponential averaging. These two noise estimation techniques have been tested with non-linear spectral subtraction as a pre-processing step to a hidden Markov model (HMM) recognizer [44].

Several methods are based on temporal minima-tracking of the smoothed noisy speech power estimate $|\hat{Y}(\omega,k)|^2$, given by

$$|\hat{Y}(\omega,k)|^2 = \lambda_Y.|\hat{Y}(\omega,k-1)|^2+(1-\lambda_Y).|Y(\omega,k)|^2, 0.1\le\lambda_Y\le 0.3 \tag{2.13}$$

This equation (2.13) is similar to (2.10), but in this case, as the speech signal is non-stationary, less averaging can be performed, leading to decreased values for λ_Y compared to λ_D. This method described in [45] looks for the minimum of $|\hat{Y}(\omega,k)|^2$ within a finite window of length L. The basic assumption of this algorithm is that peaks of the power spectrum correspond to speech activity, while valleys correspond to the smoothed noisy speech power. Valleys, can therefore, be used to estimate noise. The noise power estimate $|\widehat{D}(\omega,k)|^2$ is obtained as a weighted minimum of $|\hat{Y}(\omega,k)|^2$ within a window of L samples. The window length L must be long enough to contain speech activity peaks but short enough to follow the non-stationarities of noise (0.8 to 1.4 seconds). The performance of this algorithm depends on the trade-off between the smoothing constant λ_Y and window length L. For stationary noise, the performance is similar to classical noise estimation with an ideal speech/noise detector. This method is of particular interest in the case of non-stationary noise because it allows one to update the noise estimate even during speech activity frames. A variation of this method is proposed in [46]. This method is also based on a type of temporal minima tracking of $|\hat{Y}(\omega,k)|^2$, but it is computationally more efficient. The noise estimate is obtained with the following equation:

$$|\hat{D}(\omega,k)|^2 = \begin{cases} \gamma|\hat{D}(\omega,k-1)|^2 + \left[\frac{1-\gamma}{1-\beta}\right]\left[|\hat{Y}(\omega,k)|^2 - \beta|\hat{Y}(\omega,k-1)|^2\right], & \text{if } |\hat{Y}(\omega,k)|^2 > |\hat{D}(\omega,k-1)|^2 \\ |\hat{Y}(\omega,k)|^2, & \text{otherwise} \end{cases} \tag{2.14}$$

Here, β and γ are constants which are determined experimentally. Typical values for the parameters are $\beta = 0.96$, and $\gamma = 0.998$ which lead to an adaptation period ranging from 0.2 to 0.4 seconds.

2.5. Improvements to Spectral Subtraction Method: Spectral Subtractive-Type Algorithms

In practice, the spectral subtraction method is imperfect, due to in-accurate estimation of the noise spectrum. Many improvements to the basic idea have been suggested over the years to address its limitations. These variations form a family of spectral subtractive-type algorithms. The key idea of this technique is to estimate the short-time spectral magnitude of speech by subtracting estimated noise from the noisy speech spectrum (or by multiplying the noisy spectrum with gain functions) and to combine it with the phase of the noisy speech. The forms of subtractive-type algorithms most notably, spectral over-subtraction [20], multi-band spectral subtraction [21], Wiener filtering [22], iterative spectral subtraction [23], and spectral subtraction based on perceptual properties [24] in noisy environments.

2.5.1 Spectral Over-Subtraction Algorithm

An improved version of spectral subtraction method was proposed in [20] to minimise the annoying noise. In this algorithm, the spectral subtraction method [13] uses two additional parameters, namely, over-subtraction factor and noise spectral floor parameter [20]. The algorithm is given as

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha.|\widehat{D}(\omega)|^2, & \text{if } \frac{|\widehat{D}(\omega)|^2}{|Y(\omega)|^2} < \frac{1}{\alpha+\beta} \\ \beta.|\widehat{D}(\omega)|^2, & \text{else} \end{cases} \tag{2.15}$$

with $\alpha \geq 1$ and $0 \leq \beta \ll 1$

The over-subtraction factor (α) controls the amount of noise power spectrum subtracted from the noisy speech power spectrum in each frame and spectral floor parameter (β) prevent the resultant spectrum from going below a preset minimum level rather than setting to zero (spectral floor). Put differently, the over-subtraction factor determines the balance of the amount of noise reduction and speech distortion whereas the noise spectral flooring mask the remnant noise. The over-subtraction factor depends on the a-*posteriori* segmental SNR. The over-subtraction factor can be calculated as

$$\alpha = \alpha_0 + (\mathrm{SNR} - \mathrm{SNR}_{\min})\left(\frac{\alpha_{\min} - \alpha_0}{\mathrm{SNR}_{\max} - \mathrm{SNR}_{\min}}\right), \text{ if } \mathrm{SNR}_{\min} \leq \mathrm{SNR} \leq \mathrm{SNR}_{\max} \tag{2.16}$$

where

$$\mathrm{SNR(dB)} = 10\log_{10}\left(\frac{\sum_{k=0}^{N-1}|Y(\omega)|^2}{\sum_{k=0}^{N-1}|\widehat{D}(\omega)|^2}\right) \tag{2.17}$$

Here N is the number of samples in the signal, k is the frame index. The value of $\alpha_{\min} = 1$, $\alpha_{\max} = \alpha_0$, $\mathrm{SNR}_{\min} = 0$ dB, $\mathrm{SNR}_{\max} = 20$ dB and α_0 ($\alpha_0 \approx 4$), used in (2.16), is the desired value of α at 0 dB SNR. These values are estimated by experimental trade-off results. The relation between over-subtraction factor and segmental SNR is shown in Fig. 2.3.

The implementation of SOS algorithm assumes that the noise affects the speech spectrum uniformly and the subtraction factor subtracts an over estimate of noise from noisy spectrum. Therefore, for a balance between broadband and musical tone removal, various combinations of over-subtraction factor α, and spectral floor parameter β give rise to a trade-off between the amount of remaining broadband noise and the level of perceived musical tone. For large value of β, the spectral floor is high, and a very little, if any musical tone is audible, while with small β, the broadband noise is greatly reduced, but the musical tone becomes quite annoying. Hence, the suitable value of α is set as per (2.16), and $\beta = 0.03$.

This algorithm reduces remnant musical noise, while preventing the resultant spectral components from going below a present minimum value. The level of perceived remnant noise is reduced, but background noise remains present and enhanced speech is distorted. In Fig. 2.4 the block diagram of spectral over-subtraction algorithm is shown.

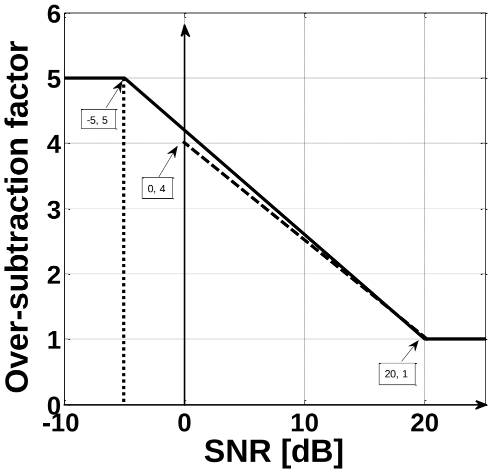


Fig. 2.3: The relation between over-subtraction factor and SNR.

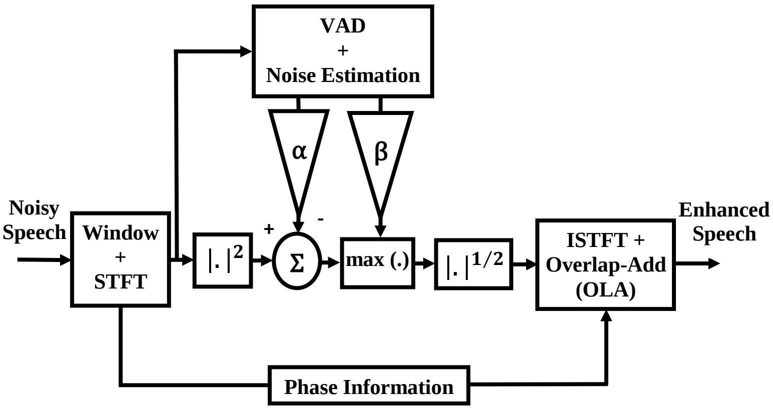


Fig. 2.4: Block diagram of spectral over-subtraction algorithm.

2.5.2 Multi-Band Spectral Subtraction Algorithm

In real-world listening environment, the noise does not affect the speech signal uniformly over the whole spectrum. Some frequencies are affected more adversely than others depending on the spectral characteristics of the noise, which eventually mean that this kind of noise is non-stationary or colored. This is best illustrated in Fig. 2.5, which shows the plot of the estimated segmental SNR of non-overlapped uniformly spaced frequency bands {60 Hz ～ 1 kHz (Band 1), 1 kHz ～ 2 kHz (Band 2), 2 kHz ～ 3 kHz (Band 3), 3 kHz ～ 4 kHz (Band 4)} over frame number. It can be seen from the figure that the segmental SNR of the low frequency bands (Band 1) is significantly higher than the segmental SNR of the high frequency bands (Band 4) [3, 21, 47]. This phenomenon suggests that the noise signal does not affect the speech signal uniformly over the whole spectrum; therefore, subtracting a constant factor of noise spectrum over the whole frequency spectrum may remove speech also.

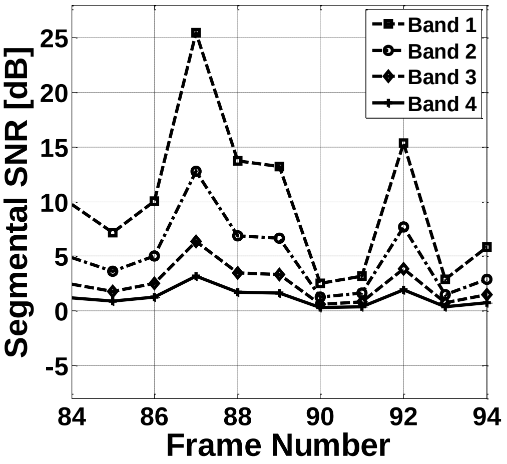


Fig. 2.5: The segmental SNR of four uniformly spaced frequency bands of degraded speech.

To take into account that the real-world noise affects the speech spectrum differently at various frequencies, a multi-band uniformly spaced frequency approach of spectral over-subtraction was presented in [21], which is the case of non-linear spectral subtraction (NSS) [41]. In this scheme, the noisy speech spectrum is divided into four uniformly spaced non-overlapping continuous frequency bands, and spectral subtraction is applied in each band, separately. The multi-band spectral subtraction algorithm re-adjusts the over-subtraction factor in each band. Therefore, the estimate of the clean speech spectrum in the i^{th} Band is obtained by

$$|\hat{S}_i(\omega)|^2 = \begin{cases} |Y_i(\omega)|^2 - \alpha_i.\delta_i.|\widehat{D}_i(\omega)|^2, & \text{if } |\hat{S}_i(\omega)|^2 > \beta.|Y_i(\omega)|^2 \\ \beta.|Y_i(\omega)|^2, & \text{else} \end{cases} \quad \omega_i < \omega < \omega_{i+1} \qquad (2.18)$$

where ω_i and ω_{i+1} are the start and end frequency bins of the i^{th} Band; α_i is the band specific over-subtraction factor of the i^{th} Band, which is the function of segmental SNR of the i^{th} frequency band (i.e. SNR_i) and provides a degree of control over the noise subtraction level in each band. The segmental SNR of the i^{th} Band (SNR_i) can be calculated as

$$\text{SNR}_i\text{ (dB)} = 10\log_{10}\left(\frac{\sum_{\omega=\omega_i}^{\omega_{i+1}}|Y_i(\omega)|^2}{\sum_{\omega=\omega_i}^{\omega_{i+1}}|\widehat{D}_i(\omega)|^2}\right) \qquad (2.19)$$

The band specific over-subtraction can be calculated using Fig. 2.3 as

$$\alpha_i = \begin{cases} \alpha_{\max}, & \text{if SNR}_i \le \text{SNR}_{\min} \\ \alpha_{\max} + (\text{SNR}_i - \text{SNR}_{\min})\left(\frac{\alpha_{\min}-\alpha_{\max}}{\text{SNR}_{\max}-\text{SNR}_{\min}}\right), & \text{if SNR}_{\min} \le \text{SNR}_i \le \text{SNR}_{\max} \\ \alpha_{\min}, & \text{if SNR}_i \ge \text{SNR}_{\max} \end{cases} \qquad (2.20)$$

Here $\alpha_{\min} = 1, \alpha_{\max} = 5, \text{SNR}_{\min} = -5\text{ dB}, \text{SNR}_{\max} = 20\text{ dB}$. These values are estimated by experimental trade-off results.

The δ_i is an additional band subtraction factor that can be individually set for each frequency band to customize the noise removal process and provide an additional degree of control over the noise subtraction level in each band. The values of δ_i [21] is empirically calculated as most of the speech energy is concentrated below 1 kHz and set to

$$\delta_i = \begin{cases} 1 & , f_i \le 1\text{ kHz} \\ 2.5 & , 1\text{ kHz} < f_i \le \frac{f_s}{2} - 2\text{ kHz} \\ 1.5 & , f_i > \frac{f_s}{2} - 2\text{ kHz} \end{cases} \qquad (2.21)$$

Here f_i is the upper bound frequency of the i^{th} Band and f_s is the sampling frequency. The motivation for using smaller values of δ_i for the low frequency bands is to minimize speech distortion, since most of the speech energy is present in the lower frequencies. Both factors, α_i and δ_i can be adjusted for each band for different speech conditions to get better speech quality.

As the real-world noise is highly random in nature, improvement in the multi-band spectral subtraction (MBSS) algorithm for reduction of WGN is necessary. But the performance of MBSS algorithm has been found to be better than other subtractive-type algorithms [13, 20] and has been demonstrated in [47-50]. In Fig. 2.6 the block diagram of multi-band spectral subtraction algorithm [50] is shown.

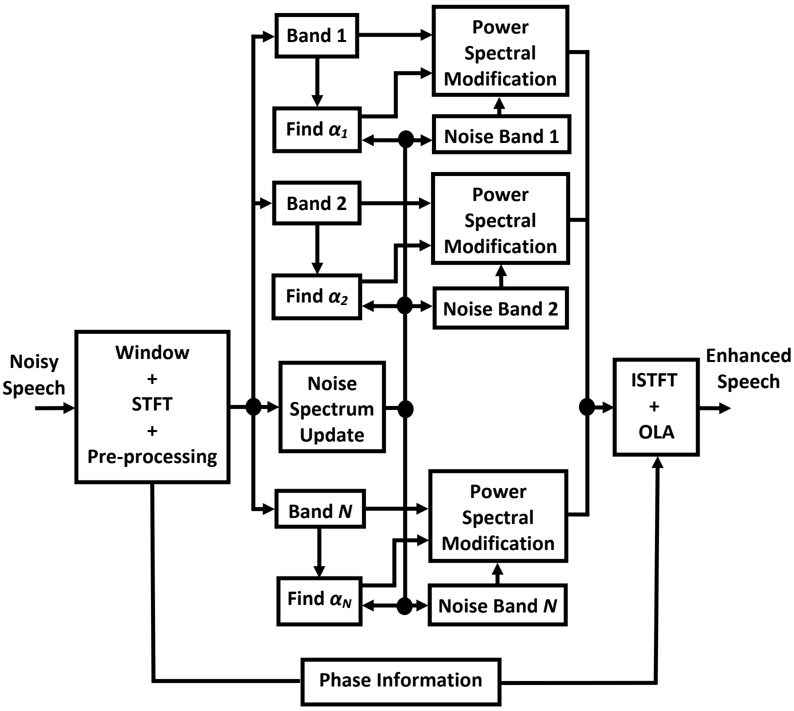


Fig.2.6: Block diagram of multi-band spectral subtraction algorithm.

2.5.3 Wiener Filtering

The spectral subtraction method can also be viewed as a filtering operation, by manipulating (2.8) such that, it can be expressed as the product of noisy speech signal spectrum and the frequency response of a spectral subtraction filter (SSF) [13] as

$$
\begin{aligned}
|\hat{S}(\omega)|^b &= |Y(\omega)|^b - |\widehat{D}(\omega)|^b \\
&= H(\omega).|Y(\omega)|^b \qquad (2.22)
\end{aligned}
$$

where $H(\omega)$, the frequency response of the spectral subtraction filter, is defined as

$$H(\omega) = \left[1 - \frac{|\widehat{D}(\omega)|^b}{|Y(\omega)|^b}\right]$$

$$= \left[\frac{|Y(\omega)|^b - |\widehat{D}(\omega)|^b}{|Y(\omega)|^b}\right]$$

$$= \left[\frac{|Y(\omega)|^b - |\widehat{D}(\omega)|^b}{|Y(\omega)|^b}\right] \qquad (2.23)$$

The spectral subtraction filter $H(\omega)$ is a zero phase filter, with its magnitude response in the range of $0 \leq H(\omega) \leq 1$. The filter acts as a SNR dependent attenuator. The attenuation in each frequency increases with the decreasing SNR, and vice-versa.

The Wiener filter (WF) is a frequency domain filter and was suggested as an improvement to the spectral subtraction [38]. Rather than direct subtraction (as in case of the spectral subtraction method), a wiener gain function is calculated, and then multiplied by the noisy spectrum to attenuate noise component frequencies. Put differently, Wiener filtering is replaces the direct subtraction with a mathematically optimal estimate for the signal spectrum in a minimum mean squared error (MMSE) sense between clean speech and the estimated speech $\mathrm{E}\left\{(S(\omega) - \hat{S}(\omega))^2\right\}$ [22].

Here, it is assumed that the speech and the noise obey normal distribution and do not correlate. The implementation of a WF requires the power spectrum of the signal and the noise. However, SSF can be used as a substitute for the WF when the signal spectrum is not available. The gain of the WF [22], $H_{\text{wiener}}(\omega)$, can be expressed in terms of the power spectrum of clean speech $P_s\,(\omega)$ and the power spectrum of noisy speech $P_y\,(\omega\,)$. But power spectrum of clean speech is not known, the power spectrum of the noisy speech signal $P_y\,(\omega)$ is used instead as

$$H_{\text{wiener}}(\omega) = \frac{P_s\,(\omega\,)}{P_y(\omega)}$$

$$= \frac{P_s\,(\omega\,)}{P_s\,(\omega\,) + P_d(\omega\,)}$$

$$= \frac{P_y\,(\omega\,) - P_d\,(\omega\,)}{P_y\,(\omega\,)} \qquad (2.24)$$

The weakness of the WF is that the fixed frequency response at all frequencies and the requirement to estimate the power spectral density of the clean signal and noise prior to filtering.

Therefore, non-causal WF cannot be applied directly to estimate the clean speech since speech cannot be assumed to be stationary. Therefore, an adaptive WF implementation can be used to approximate (2.24) as

$$H_{\text{A.wiener}}(\omega) = \frac{|Y(\omega)|^2 - |\widehat{D}(\omega)|^2}{|Y(\omega)|^2} \tag{2.25}$$

$$|\hat{S}(\omega)|^2 = H_{\text{A.wiener}}(\omega). |Y(\omega)|^2 \tag{2.26}$$

$H_{\text{A.wiener}}(\omega)$ attenuates each frequency component by a certain amount depending on the power of the noise at the frequency.

From (2.25), if $|\widehat{D}(\omega)|^2 = 0$, then $H_{\text{A.wiener}}(\omega) = 1$ and no attenuation takes place, i.e. there is no noise component at the frequency ω, whereas if $|\widehat{D}(\omega)|^2 = |Y(\omega)|^2$, then $H_{\text{A.wiener}}(\omega) = 0$. Therefore, the frequency component is completely nulled. All other values of $H_{\text{A.wiener}}(\omega)$ scale the power of the signal by an appropriate amount.

On comparing $H(\omega)$ and $H_{\text{A.wiener}}(\omega)$ from (2.23) and (2.25), it can be observed that the WF is based on the ensemble-average spectra of the signal and noise, whereas the SSF ($b = 2$) uses the instantaneous spectra for noise signal and the running average (time-averaged spectra) of the noise. In WF theory, the averaging operations are taken across the ensemble of different realization of the signal and noise processes. In spectral subtraction, we have access only to single realization of the process.

2.5.4 Iterative Spectral Subtraction Algorithm

An iterative spectral subtraction (ISS) algorithm is proposed in [23], motivated from WF [22, 38], to suppress the remnant musical noise. In this algorithm, the output of the enhanced speech is used as the input signal for the next iteration process. As after the spectral subtraction process, the type of the additive noise is changed to the remnant musical noise and the output signal is used as the input signal of the next iteration process. The remnant noise is re-estimated and this new estimated noise, furthermore, is used to process the next spectral subtraction. Therefore, an enhanced output speech signal can be obtained, and the iteration process goes on. If we regard the process of noise estimate and the spectral subtraction as a filter, the filtered output is used not only for designing the filter but also as the input of the next iteration process.

The iteration number is the most important factor of this algorithm which directly influence the performance of speech enhancement. The larger iteration number corresponds to better speech enhancement with the less remnant noise [51-54]. In Fig. 2.7 the block diagram of iterative spectral subtraction algorithm is shown.

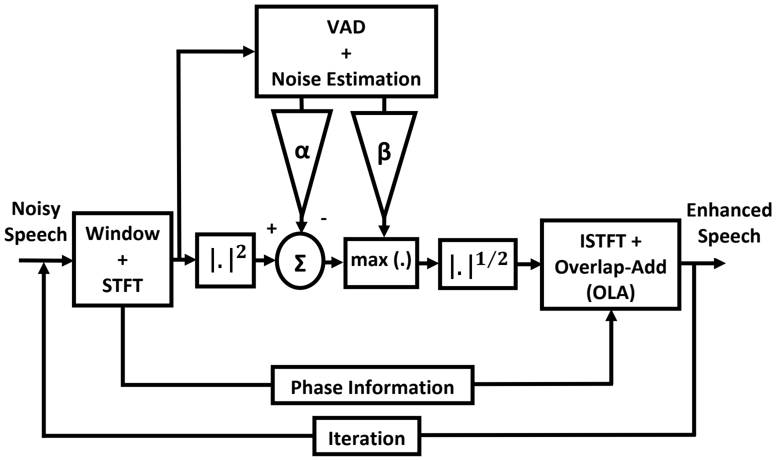


Fig. 2.7: Block diagram of iterative spectral subtraction algorithm.

2.5.5 Spectral Subtraction based on Perceptual Properties

The main weakness of spectral over-subtraction algorithm [20] is that it uses the fixed value of subtraction parameters that are unable to adapt the varying noise levels and noise characteristics. However, the optimization of the parameters is not an easy task, because the spectrum of most of the noise which is added in speech is not flat but non-stationary or colored. An example of adaptation is multi-band spectral subtraction algorithm (non-linear algorithm), this scheme adopts the subtractive parameters α and β in time and frequency based on the segmental signal-to-noise ratio, leading to improved results but remnant noise is not suppressed completely at low SNR's [21]. Put differently, in lower SNR conditions, it’s difficult to find the best trade-off between the amount of noise reduction, the speech distortion and the level of remnant noise in a perceptual sense. Therefore, the selection of appropriate value of subtractive parameters is the major task in subtractive-type algorithms for enhancement of noisy speech.

The concept of masking threshold of human auditory system is explored in [24], to reduce the annoying remnant noise below the noise masking threshold of clean speech signal and to make less speech distortion. When two signals are close in time or frequency, one is rendered completely or partially inaudible by the other. This is known as auditory masking. In this approach, the subtraction parameters are adapted based on the noise masking threshold of HAS to achieve a good trade-off between the remnant noise, and speech distortion. If the masking threshold is high, the remnant noise will be masked naturally and it will not be audible. In this case, the subtraction parameters have their minimum values, thereby reducing speech distortion. However, if the masking threshold is low, the remnant noise is not masked. In this case, it is necessary to increase the values of subtraction parameters. The adaptation of subtraction parameters is done according to the relations

$$\alpha = \begin{cases} \alpha_{\max}\,, & \text{if } T(\omega) = T(\omega)_{\min} \\ \alpha_{\min}\,, & \text{if } T(\omega) = T(\omega)_{\max} \\ \alpha_{\max}\left(\frac{T(\omega)_{\max}-T(\omega)}{T(\omega)_{\max}-T(\omega)_{\min}}\right) + \alpha_{\min}\left(\frac{T(\omega)-T(\omega)_{\min}}{T(\omega)_{\max}-T(\omega)_{\min}}\right), & \text{if } T(\omega)\epsilon[T(\omega)_{\min}, T(\omega)_{\max}] \end{cases} \quad (2.27)$$

$$\beta = \begin{cases} \beta_{\max}\,, & \text{if } T(\omega) = T(\omega)_{\min} \\ \beta_{\min}\,, & \text{if } T(\omega) = T(\omega)_{\max} \\ \beta_{\max}\left(\frac{T(\omega)_{\max}-T(\omega)}{T(\omega)_{\max}-T(\omega)_{\min}}\right) + \beta_{\min}\left(\frac{T(\omega)-T(\omega)_{\min}}{T(\omega)_{\max}-T(\omega)_{\min}}\right), & \text{if } T(\omega)\epsilon[T(\omega)_{\min}, T(\omega)_{\max}] \end{cases} \quad (2.28)$$

Here $\alpha_{\max}, \alpha_{\min}, \beta_{\max}, \beta_{\min}$ and $T(\omega)_{\max}, T(\omega)_{\min}$ are the maximal and minimal values of $\alpha,\ \beta$ and updated masking threshold $T(\omega)$, respectively [24, 55]. It can be seen from (2.27) and (2.28) that α, β achieves the maximal and the minimal values when $T(\omega)$ equals its minimal and maximal values. The noise masking threshold can be calculated from the enhanced speech as the method proposed by [56]. In Fig. 2.8, the block diagram of spectral subtraction based on perceptual properties is shown.

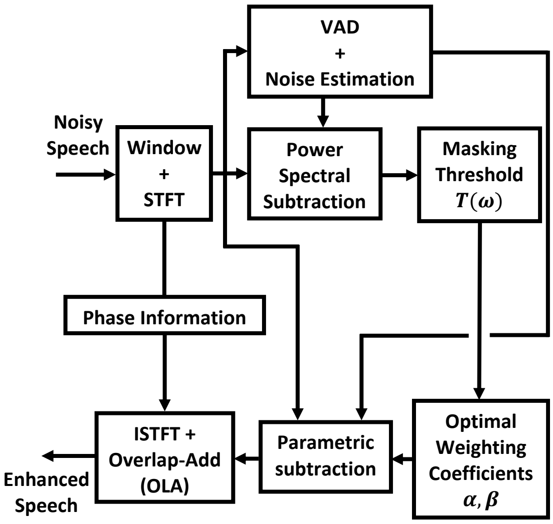

VAD
+
Noise Estimation
Noisy Speech
Window
+
STFT
Power Spectral Subtraction
Masking Threshold $T(\omega)$
Phase Information
ISTFT + Overlap-Add (OLA)
Parametric subtraction
Optimal Weighting Coefficients α, β
Enhanced Speech

Fig. 2.8: Block diagram of spectral subtraction based on perceptual properties.

2.6. Implementation, Experiments Results, and Performance Evaluation

This section presents the implementation, experimental results and the comparative study of performance evaluation of spectral subtractive-type algorithms described in this chapter. For simulations, we employ MATLAB software as a programming environment as it offers many advantages. It contains a variety of signal processing and statistical tools, which help users in generating a variety of signals and plotting them. MATLAB excels at numerical computations, especially when dealing with vectors or matrices of data.

The clean speech and noisy speech are taken from NOIZEUS speech corpus [76]. The NOIZEUS is a publicly available database often used for benchmark experiments. The NOIZEUS corpus composed of 30 phonetically balanced sentences pronounced by six speakers (three male and three female) in English language. The corpus is sampled at 8 kHz, quantized linearly using 16 bits resolution and filtered to simulate receiving frequency characteristics of telephone handsets. Noise signals have different time-frequency distributions and a different impact on speech. For that reason, the NOIZEUS corpus comes with various non-stationary noises at different levels of SNRs. The non-stationary noises are car, train, restaurant, babble, airport, street, and exhibition.

In our evaluation, we have used the speech degraded by car noise (the car noise energy is concentrated in the low frequencies and its spectrum show an exponential decay when frequency increases) at global SNR levels of 0 dB to 15 dB in steps of 5 dB. We also generate a corresponding stimulus set degraded by additive white Gaussian noise (WGN) (stationary) at four SNR levels: 0 dB, 5 dB, 10 dB, and 15 dB. The performance of the spectral subtractive-type algorithms is tested on such noisy speech samples.

In our experiments, the noise samples used are of zero mean and the energy of the noisy speech samples are normalized to unity. The frame size is chosen to be 256 samples (32 ms — a time frame), with 50% overlapping. The sinusoidal Hamming window with size 256 samples is applied on each frame before it is enhanced individually. The windowed speech frame is then analyzed using the fast Fourier transform (FFT). We employ FFT length of 256 samples. The noise estimate is updated during the silence frames by using averaging (2.10). The final enhanced speech is reconstructed from the enhanced frames using the weighted overlap-adds (OLA) methods.

For SOS algorithm, the value of α is taken same as used in Fig. 2.2 [20] and β is kept fixed at 0.03. For MBSS approach four uniformly spaced frequency bands is used with $\beta = 0.03$ and the value of δ_i is set as per (2.21). For WF, the value of smoothing constant is taken as 0.99. For ISS algorithm, the iteration time is taken as 2-3 and for SSPP algorithm the value of $\alpha_{\max} = 6, \alpha_{\min} = 1, \beta_{\min} = 0,$ and $\beta_{\max} = 0.02$ [24].

The amount of noise reduction is generally measured with the SNR improvement, given by the difference between input SNR and output SNR. The following equation is computed for evaluation of SNR results of enhanced speech signals:

$$\mathrm{SNR}_{\mathrm{imp}} = \mathrm{SNR}_{\mathrm{output}} - \mathrm{SNR}_{\mathrm{input}} \tag{2.29}$$

The $\mathrm{SNR}_{\mathrm{input}}$ is the global SNR value of the input speech signal standing for the amount of the additive noise, the $\mathrm{SNR}_{\mathrm{output}}$ is the global SNR value of the output enhanced speech signal standing for the speech enhancement scheme and the submission is performed over the signal length. The SNR can be calculated as follows:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{n=1}^{L} s^2(n)}{\sum_{n=1}^{L}\{s(n)-\hat{s}(n)\}^2}\right) \tag{2.30}$$

where $s(n)$ is the clean speech signal, $\hat{s}(n)$ is the enhanced speech reproduced by a speech processing system, n is the sample index, and L is the number of samples in both speech signals. The summation is performed over the signal length.

The SNR improvement is the performance evaluation for calculating the amount of noise reduction in the background noise level conditions. The obtained value of SNR improvement for WGN for different enhancement algorithms is presented in Fig. 2.9. The best noise reduction is obtained in case of SSPP algorithm. The main drawback of the SNR is that it has a poor correlation with subjective quality assessment results. Therefore, the SNR of enhanced speech is not a sufficient objective measure of speech quality.

Normally, spectral subtractive-type speech enhancement algorithms generate two main undesirable effects, i.e., remnant musical noise and speech distortion. These two effects can be annoying to a human listener, and causes listeners fatigue. However, they are difficult to quantify. Therefore, it is important to analyze the time-frequency distribution of the enhanced speech, in particular the structure of its remnant noise. The speech spectrogram is a good tool to do this work, because it can give more accurate information about remnant musical noise and speech distortion than the corresponding time domain waveforms. For comparative purpose, Fig. 2.10 shows the plot of spectrogram of the clean speech signal, noisy speech (degraded by WGN) and speech enhanced by the different spectral subtractive-type algorithms, namely, BSS, SOS, MBSS, WF, ISS, and SSPP. Fig. 2.11 shows the spectrogram of enhanced speech in case of car noise.

Fig. 2.10 (iii) presents the enhanced speech obtained with basic spectral subtraction (i.e. magnitude spectral subtraction) algorithm with no remnant noise reduction. The remnant noise level is very important and its musical structure can be observed. This shows that this basic method cannot be used at very low SNR without any improvement.

The spectral over-subtraction (i.e. modified spectral subtraction) is a well-known method for removing remnant noise, as explained in Section 2.4.1.1. This parametric formulation is given in (2.15), and allows us to vary the parameters α and β. However, the choice of parameters is not easy and depends on the noise level and noise type. Fig. 2.10 (iv) shows an enhanced speech spectrogram obtained with this algorithm. The remnant noise is reduced compared to BSS.

Fig. 2.10 (v)-(viii) shows an enhanced speech spectrogram obtained with algorithms MBSS, ISS, WF, and SSPP algorithm. From the spectrogram, we can easily observe that the MBSS, ISS, and Wiener filtering have a very small amount of remnant noise and spectral subtraction based on perceptual properties has a better performance comparative to other algorithms for speech enhancement. Wiener filtering results in a smaller amount of remnant noise, but this noise has musical structure and speech regions, especially fricative consonants, are also attenuated. This type of BSS can result in speech distortion. Also, in case of car noise the BSS, SOS, ISS, and WF results are weak comparative to MBSS and SSPP.

The best results were obtained with spectral subtraction with perceptual properties. In case of this type of subtractive-type algorithm small amount of remnant noise is remaining, but this noise has a perceptually white quality and distortion remains acceptable. Informal listening tests also indicated that the enhanced speech with SSPP algorithm is more pleasant, the remnant noise is better reduced, and with minimal, if any, speech distortion.

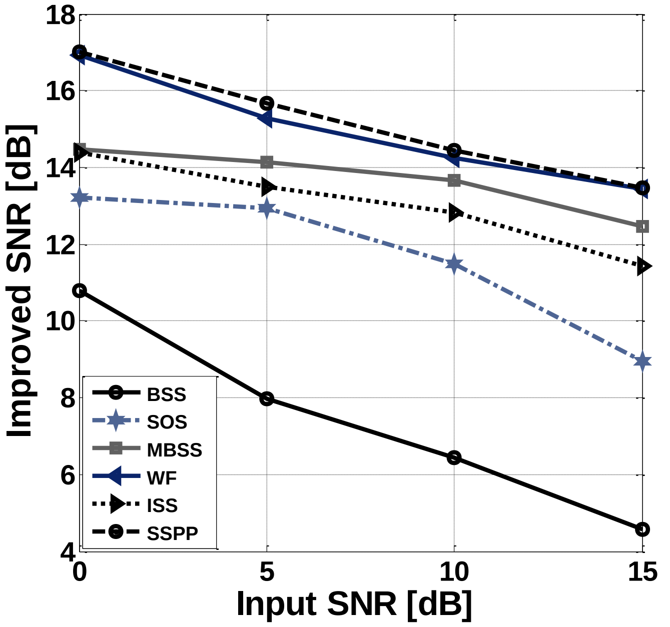
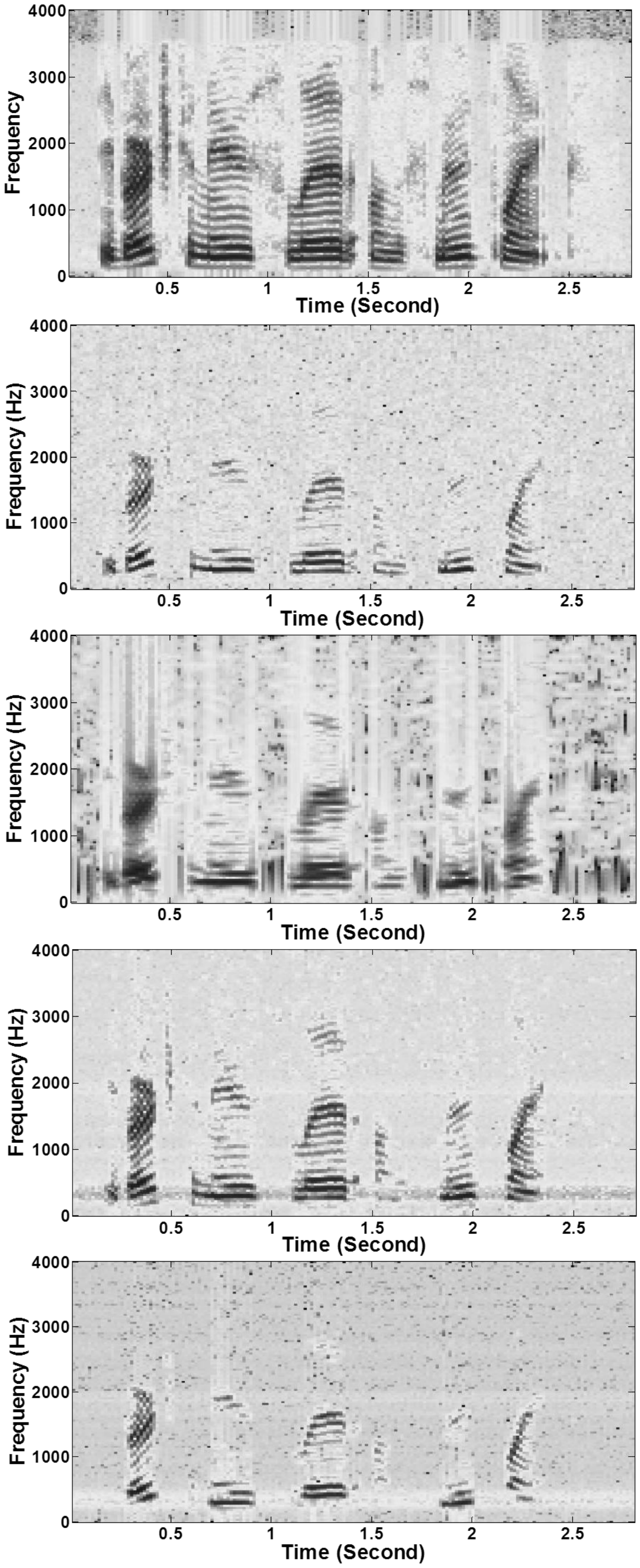


Fig. 2.9: Improved SNRs of different subtractive-type algorithms for WGN.


4000
3000
2000
1000
0
Frequency
0.5
1
1.5
2
2.5
Time (Second)
4000
3000
2000
1000
0
Frequency (Hz)
0.5
1
1.5
2
2.5
Time (Second)
4000
3000
2000
1000
0
Frequency (Hz)
0.5
1
1.5
2
2.5
Time (Second)
4000
3000
2000
1000
0
Frequency (Hz)
0.5
1
1.5
2
2.5
Time (Second)
4000
3000
2000
1000
0
Frequency (Hz)
0.5
1
1.5
2
2.5
Time (Second)

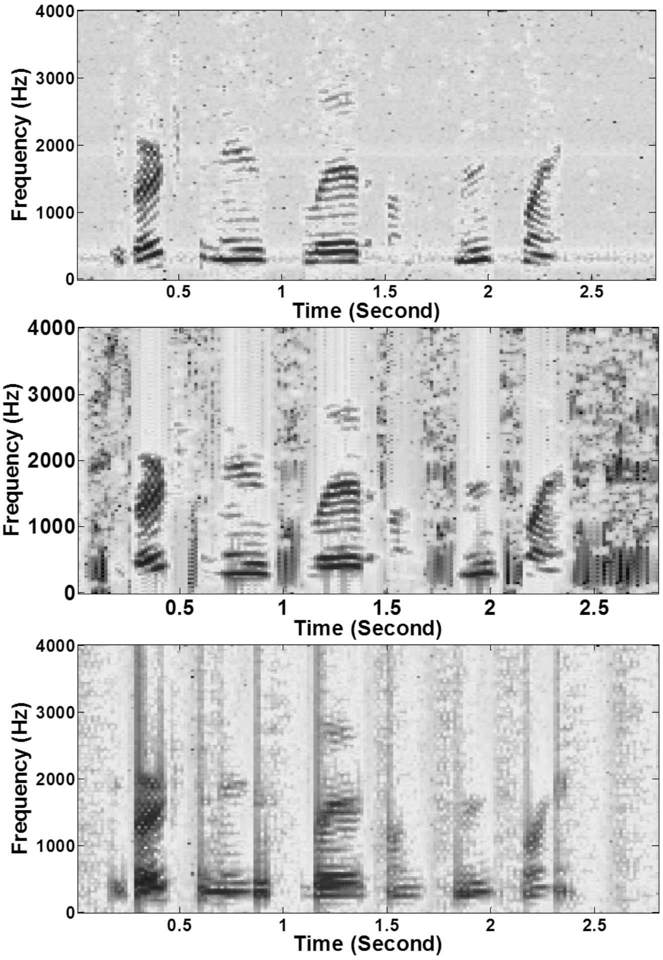
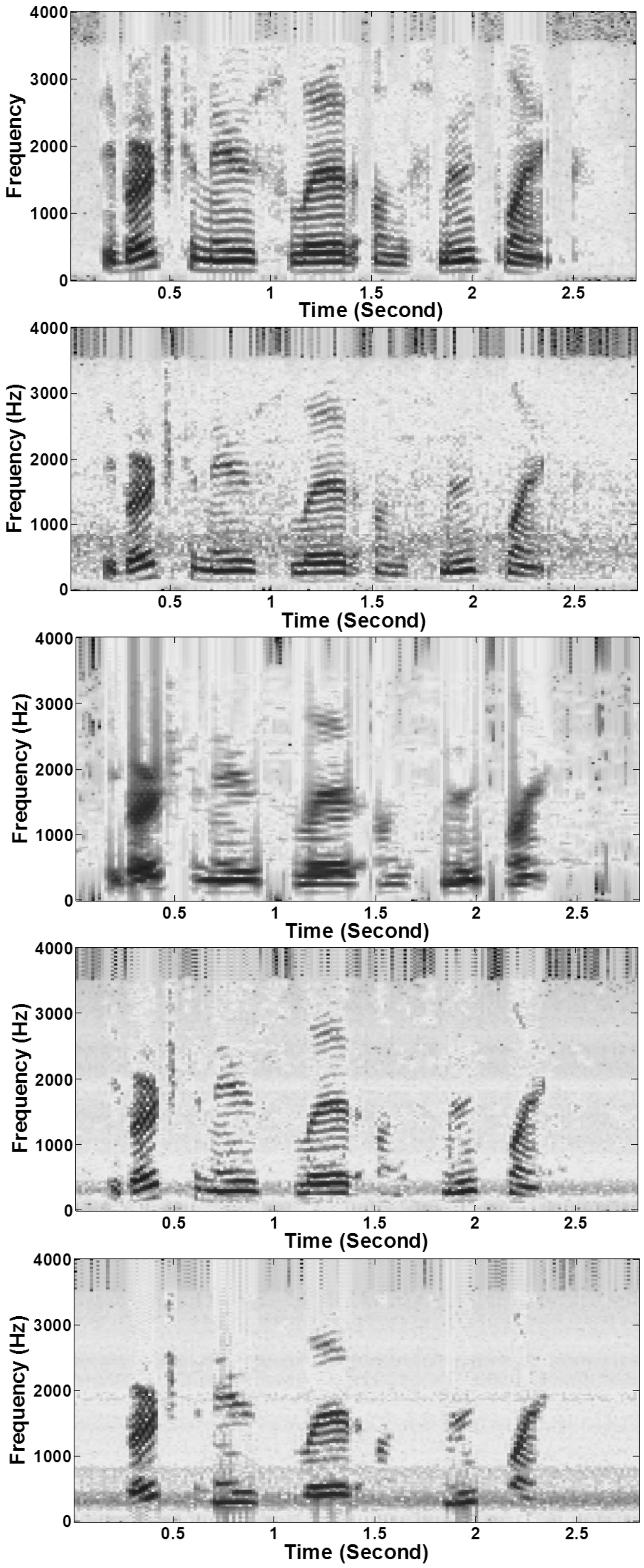


Fig. 2.10: Speech spectrogram (From top to bottom): (i) clean speech, (ii) noisy speech at SNR = 15dB, (iii) – (viii) speech enhanced by different subtractive-type algorithms; (iii) speech enhanced by BSS, (iv) speech enhanced by SOS, (v) speech enhanced by MBBS, (vi) speech enhanced by ISS, (vii) speech enhanced by WF, and (viii) speech enhanced by SSPP.


4000
3000
2000
1000
0
Frequency
0.5
1
1.5
2
2.5
Time (Second)
4000
3000
2000
1000
0
Frequency (Hz)
0.5
1
1.5
2
2.5
Time (Second)
4000
3000
2000
1000
0
Frequency (Hz)
0.5
1
1.5
2
2.5
Time (Second)
4000
3000
2000
1000
0
Frequency (Hz)
0.5
1
1.5
2
2.5
Time (Second)
4000
3000
2000
1000
0
Frequency (Hz)
0.5
1
1.5
2
2.5
Time (Second)

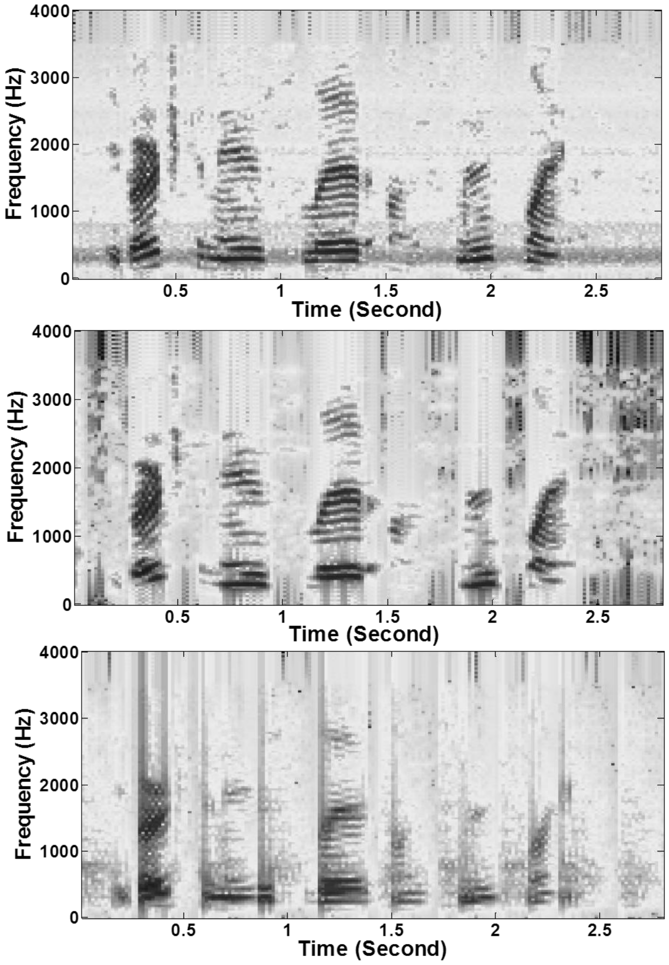


Fig. 2.11: Speech spectrogram (From top to bottom): (i) clean speech, (ii) degraded speech (car noise at SNR = 15dB), (iii) – (viii) speech enhanced by different subtractive-type algorithms; (iii) speech enhanced by BSS, (iv) speech enhanced by SOS, (v) speech enhanced by MBBS, (vi) speech enhanced by ISS, (vii) speech enhanced by WF, and (viii) speech enhanced by SSPP.

2.7. Summary

In this chapter, a comparative and simulation study of different derivatives of spectral subtractive-type algorithms is presented with a unified view of the single channel speech enhancement algorithms in the frequency domain. In particular, algorithms based on short-time Fourier transforms are examined. The limitations of spectral subtraction are briefly discussed. The artifact introduced by spectral subtraction method and the way the conventional spectral subtraction method is modified to counter these artifacts is described.

The performance evaluation of these approaches was carried out using objective measure of speech quality and informal subjective listening tests. Classical spectral subtraction method mostly results in audible remnant noise, which decreases speech intelligibility. The most progressive algorithm of speech enhancement is the spectral subtraction based on perceptual properties (SSPP), i.e. masking properties of psychoacoustic model of human auditory system. This speech enhancement algorithm takes advantage of how people perceive the frequencies instead of just working with SNR. It results in appropriate remnant noise suppression and acceptable degree of speech distortion, introduced during the enhancement process. In terms of SNR improvement, the algorithm spectral subtraction based on perceptual properties show the best noise reduction/speech enhancement results in comparison to other algorithms such as, Wiener filtering, multi-band spectral subtraction, iterative spectral subtraction. This observation is also noticed using spectrogram results.

Moreover, it is evident from the informal subjective evaluation (listening) tests that the SSPP algorithm suffers from some artifacts in the enhanced speech. Therefore, a strong noise estimation approach is needed for improving the performance of speech enhancement system.

The spectral subtractive-type algorithms presented in this chapter are the base for the development of the proposed speech enhancement algorithms in Chapter 4, Chapter 5, and Chapter 6. The next chapter, i.e., Chapter 3 presents the detailed study on stationary wavelet packet transform which is used in Chapter 6 for developing perceptually motivated stationary WPT.

Chapter 3

Stationary Wavelet Packet Transform

3.1. Introduction

The study conducted in this chapter forms the basis of the speech enhancement algorithm proposed in Chapter 6 of this thesis. This chapter begins with a brief introduction of square-integrable function in wavelet theory, and then the details of discrete wavelet transform (DWT) are presented. Next, the wavelet packet transforms (WPT) and its filterbank structure based implementation is described. Finally, this chapter describes the stationary wavelet packet transform (SWPT) and the advantage of SWPT over DWT and WPT is discussed. The stationary WPT is utilized for the time-frequency decomposition of the noisy speech signal for the perceptually motivated multi-band improved spectral over-subtraction based speech enhancement algorithm as presented in Chapter 6.

3.2. Wavelet Theory

Let us consider the space $\mathbb{L}^2(\mathbb{R})$ of the square-integrable functions in $\mathbb{R}$, where $\mathbb{R}$ is the set of real numbers. The condition for a function $s(t)$ belonging to $\mathbb{L}^2(\mathbb{R})$ is

$$\begin{aligned} \|s\|^2 &= \mathbb{E}[|s(t)|^2] \\ &= \int_{-\infty}^{+\infty} |s(t)|^2 dt < \infty \end{aligned} \tag{3.1}$$

The reason for limiting the discussion to $\mathbb{L}^2(\mathbb{R})$ space is that, in real-world the signals which we deal with have finite energy.

Let the wavelet function represent an orthonormal basis to the space of $\mathbb{L}^2(\mathbb{R})$ such that $\mathbb{L}^2(\mathbb{R}) = \overline{\text{span}}\{\psi_{a,b}(t); a\epsilon\mathbb{R}^+, b \in \mathbb{R}\}$ be a family of functions defined as scales (i.e. dilation) and translations (i.e. shifts) of a single function $\psi(t)$ as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right), a\epsilon\mathbb{R}^+, b \in \mathbb{R}\ (a \neq 0) \qquad (3.2)$$

where t is time, b is the translation parameter, and a is the scale parameter. The parameter t provides the time location of the window and it varies as the window is shifted through the signal, while parameter a controls the amount of stretching or compressing of the mother wavelet $\psi(t)$. A large value of parameter a stretches the basic wavelet function and allows the analysis of low-frequency components of the signal. A small value of a gives a contracted version of the basic wavelet and then allows the analysis of high frequency components.

Here a is restricted to $\mathbb{R}^+$; which is natural since a can be interpreted as a reciprocal of frequency i.e. $(a = 1/f)$, and f represents the frequency. Normalization by $\frac{1}{\sqrt{|a|}}$ ensures that $\|\psi_{a,b}(t)\|$ is independent of a and b. Wavelet functions are usually normalized to 'have unit energy', i.e., $\|\psi_{a,b}(t)\| = 1$.

The function ψ (is the transforming function, called the wavelet function or mother wavelet function) is assumed to satisfy the admissibility condition,

$$C_\psi = \int \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty \qquad (3.3)$$

where $\Psi(\omega) = \int \psi(t)\, e^{-j\omega t} dt$ is the Fourier transformation of $\psi(t)$. The admissibility condition (3.3) implies

$$0 = \Psi(0) = \int \psi(t) dt \qquad (3.4)$$

which means that $\psi(t)$ is a zero mean function in the time domain and resembles a band-pass filter in the frequency domain. Also, if $\int \psi(t)dt = 0$ and $\int (1 + |t|^\alpha)|\psi(t)|\, dt < \infty$ for some $\alpha > 0$, then $C_\psi < \infty$. For any signal $s(t), s(t)\epsilon\, \mathbb{L}^2(\mathbb{R})$, the continuous wavelet transformation (CWT) is defined as a function of two variables [25, 26]

$$\text{CWT}_s(a, b) = \langle s(t), \psi_{a,b}(t)\rangle$$

$$= \int_{-\infty}^{\infty} s(t) \,.\, \psi_{a,b}^{*}(t) \, dt \tag{3.5}$$

where $*$ denotes complex conjugation and $\langle .. , .. \rangle$ is the $\mathbb{L}^2(\mathbb{R})$ inner product operation. Throughout this discussion, wavelets will be assumed to be real, and so $\psi(t) = \psi^{*}(t)$.

3.3. Discrete Wavelet Transform

The continuous wavelet transformation is a function of two variables and redundant. To minimize the transformation one can select discrete values of a and b and still have a transformation that is invertible. However, sampling that preserves all information about the decomposed function cannot be coarser than the critical sampling. The critical sampling, shown in Fig. 3.1, defined by

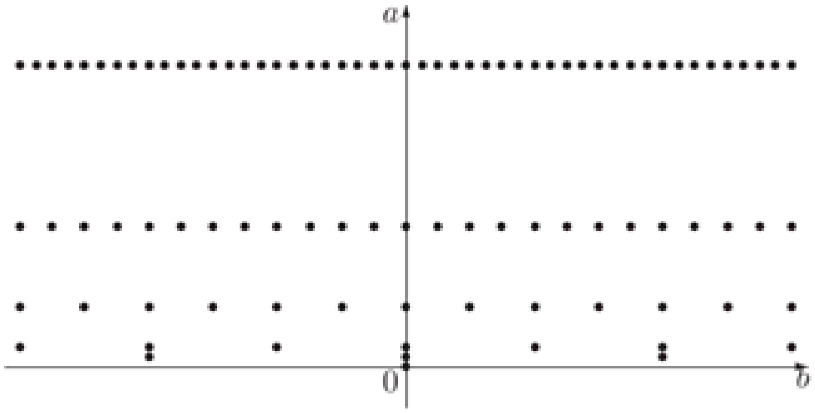


Fig. 3.1: Critical sampling in $\mathbb{R} \times \mathbb{R}^{+}$ half plane $(a = 2^{-j}$ and $b = k2^{-j})$.

$$a = 2^{-j},\ b = k2^{-j},\ j, k \in \mathbb{Z} \tag{3.6}$$

which produces the minimal basis. Any coarser sampling will not give a unique inverse transformation; i.e., the original function will not be uniquely recoverable. Moreover, under mild conditions on the wavelet function ψ, such sampling produces an orthogonal basis

$$\{\psi_{j,k}(t) = 2^{j/2}\, \psi(2^{j} t - k), j, k \in \mathbb{Z}\} \tag{3.7}$$

where $\mathbb{Z}$ denotes the set of integers. From (3.7) it is clear that the scaling function can be expressed as a linear combination of the half scale scaling function and its shifted versions which are orthogonal to each other. In this case, the space spanned by the scaling function with larger scale is included in the space spanned by the scaling function with smaller scale. In other words, the space spanned by the scaling function with larger scale is subspace of the space spanned by the scaling function with smaller scale.

There are other discretization choices. For example, selecting $a = 2^{-j}, b = k$ will lead to non-decimated (or stationary) wavelets. For more general sampling, given by

$$a = a_0^{-j}, a_0 > 1 \,, b = kb_0 a_0^{-j},\ b_0 \neq 0,\ \ j,k \in \mathbb{Z} \tag{3.8}$$

Numerically, stable reconstructions are possible if the system $\psi_{j,k},\ \ j,k \in \mathbb{Z}$ constitutes a frame. From (3.2)

$$\begin{aligned}\psi_{j,k}(t) &= a_0^{j/2}\psi\left(\frac{t - kb_0 a_0^{-j}}{a_0^{-j}}\right)\\ &= a_0^{j/2}\psi\left(a_0^{j}t - k\,b_0\right)\end{aligned} \tag{3.9}$$

Next, we consider wavelet transformations (wavelet series expansions) for values of a and b given by (3.8). An elegant theoretical framework for critically sampled wavelet transformation is Mallat's multi-resolution analysis (MRA) [55].

3.4. Multi-Resolution (or Multi-Scale) Analysis

The orthogonal wavelet expansion can also be seen as a multi-resolution formulation. According to Mallet, and Burrus *et* al. in [57, 58], there are two main components in the multi-resolution formulation of wavelet analysis, namely scaling and wavelet functions. The scaling function can be defined as

$$\phi_k(t) = \phi(t - k), \qquad k\epsilon\mathbb{Z} \text{ and } \phi\epsilon\mathbb{L}^2(\mathbb{R}) \tag{3.10}$$

where k is the discrete step translation**.** The subspace of $\mathbb{L}^2(\mathbb{R})$ spanned by these translates $\phi(t-k)$ is defined as

$$V_0 = \overline{\text{Span}_k\{\phi_k(t)\}} \tag{3.11}$$

For all integers $\mathbb{Z}$, $-\infty < k < \infty$. The over-line denotes the closure of a space. Then, a two-dimensional family of functions is generated from the basic scaling function by dilation and translation as

$$\phi_{j,k}(t) = 2^{j/2}\phi\left(2^j t - k\right) \qquad j,k\epsilon\mathbb{Z} \text{ and } \phi\epsilon\mathbb{L}^2(\mathbb{R}) \tag{3.12}$$

whose span over k is

$$V_j = \overline{\underset{k}{\text{Span}}\{\phi_k(2^J t)\}}$$

$$= \overline{\text{Span}_k\{\phi_{J,k}(t)\}} \tag{3.13}$$

for all integers $j, k \in \mathbb{Z}$. The subspace V_j, in (3.13) has nesting property [57] such that

$$\{0\} \subset \cdots \subset V_{-1} \subset V_{-1} \subset V_1 \subset \cdots V_j \subset \cdots \subset \{\mathbb{L}^2(\mathbb{R})\} \tag{3.14}$$

It is obvious that as j goes to infinity V_j enlarges to cover all energy signals $\mathbb{L}^2(\mathbb{R})$. On the other hand, as j goes to minus infinity V_j shrinks down to cover only the zero signal [57].

The difference between space spanned by scaling function and its half scale version is expressed as the orthogonal complement. This orthogonal complement is spanned by the corresponding wavelet function. This means, if we have a certain scaling function with particular scale, the space spanned by that scaling function can be decomposed into a subspace and its orthogonal complement. The subspace is spanned by the scaling function with double scale of the previous scaling function while the orthogonal complement is spanned by the corresponding wavelet function. Therefore we can define the space spanned by wavelet W_j as

$$V_{j+1} = V_j \oplus W_j \tag{3.15}$$

If the V space is further decomposed, then the following is obtained

$$V_{j+1} = \left(V_{j-1} \oplus W_{j-1}\right) \oplus W_j$$

$$= \left(V_{j-2} \oplus W_{j-2}\right) \oplus W_{j-1} \oplus W_j$$

$$= \left(V_{j-3} \oplus W_{j-3}\right) \oplus W_{j-2} \oplus W_{j-1} \oplus W_j$$

$$= \left(V_{j-3} \oplus W_{j-3}\right) \oplus W_{j-2} \oplus W_{j-1} \oplus W_j$$

$$V_{j+1} = (V_0 \oplus W_0) \oplus W_1 \oplus \ldots\ldots \oplus W_{j-2} \oplus W_{j-1} \oplus W_j \tag{3.16}$$

with $V_0 \perp W_0 \perp W_1 \perp \cdots \perp W_{j-2} \perp W_{j-1} \perp W_j$

where ⊥ denotes the orthogonality operator. So, the approximation space at resolution j, V_j can be written as a sum of subspaces. These subspaces are mutually orthogonal. As a consequence of (3.16), the entire square-integrable functions can be represented as

$$\mathbb{L}^2(\mathbb{R}) = V_0 \oplus W_0 \oplus W_1 \oplus W_2 \ldots..$$
$$= V_0 \oplus \left[\sum_{j=0}^{J-1} W_j\right] \tag{3.17}$$

where V_0 is the coarse approximation, W_0 is detail at level 0, W_1 is detail at level 1, W_2 is detail at level 2, and so on. The $\oplus$ is the direct sum of orthogonal operator, which corresponds to the linear closure of two orthogonal spaces. In (3.17), W_j and V_j can be written in terms of the basis functions $\{\psi_{j,k}(t)\}$ and $\{\phi_{j,k}(t)\}$ as $W_j = \text{span}\{\psi_{j,k}(t)\}$ and $V_j = \text{span}\{\phi_{j,k}(t)\}$. In discrete wavelet transform, any signal $s(t), s(t) \epsilon \mathbb{L}^2(\mathbb{R})$, can be written as a series expansion in terms of the scaling function and wavelets function on different scales as

$$s(t) = \sum_{k=-\infty}^{\infty} a_{j_0}(k)\, \phi_{j_0,k}(t) + \sum_{j=j_0}^{J} \sum_{k=-\infty}^{\infty} d_j(k)\, \psi_{j,k}(t) \tag{3.18}$$

The first summation in (3.18) gives us with a coarse approximation to $s(t)$, which is the projection of $s(t)$ onto V_{j_0}. The second summation for each j provides finer details and is the projection of $s(t)$ onto the W_J spaces. In (3.18), the DWT coefficients $a_{j_0}(k)$ (scaling coefficients) at scale 2^{j_0} and $d_j(k)$ (detailed coefficients) at scale 2^j denote the weight of scaling function $\phi_{j_0,k}(t)$ and wavelet function $\psi_{j,k}(t)$, respectively, while j_0 defines coarsest scale spanned by the scaling function [25]. Correspondingly, the DWT coefficients $a_j(k)$ and $d_j(k)$ can be expressed as

$$a_j(k) = \langle s(t), \phi_{j,k}(t)\rangle$$
$$= \int s(t)\, \phi_{j,k}(t)\, dt \tag{3.19}$$

$$d_j(k) = \langle s(t), \psi_{j,k}(t)\rangle$$
$$= \int s(t)\, \psi_{j,k}(t)\, dt \tag{3.20}$$

The functions $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ form an orthonormal basis in $\mathbb{L}^2(\mathbb{R})$.

The basis functions $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ are the two-dimensional families of functions generated from the discrete scaling function $\phi(t) = \phi_{0,0}(t)$ and the discrete analysis wavelet $\psi(t) = \psi_{0.0}(t)$ on decomposition of signal $s(t)$, and are given by

$$\phi_{j,k}(t) = 2^{j/2}\phi\left(2^j t - k\right), \quad j,k \in \mathbb{Z}, \text{with } i > j \tag{3.21}$$

$$\psi_{j,k}(t) = 2^{j/2}\psi\left(2^j t - k\right), \quad j,k \in \mathbb{Z},\ \psi\epsilon\mathbb{L}^2(\mathbb{R}) \tag{3.22}$$

Here, parameter j denotes the dilation (i.e. scale) or the visibility in frequency, and k denotes the integer translation (i.e. shift or position) of the scaling function and wavelet functions. The basis functions $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ satisfy the following [57, 58]:

i) $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ must be orthonormal bases of V_j and W_j respectively

ii) $W_j \perp W_k$ for $j \neq k$

iii) $W_j \perp V_j$

Based on these conditions, the scaling function $\phi(t)$ and $\psi(t)$ are recursively, defined as

$$\phi(t) = \sqrt{2}\sum_{k=-\infty}^{\infty} h(k)\phi(2t - k)\,, k\epsilon\mathbb{Z} \tag{3.23}$$

$$\psi(t) = \sqrt{2}\sum_{k=-\infty}^{\infty} g(k)\phi(2t - k)\,, k\epsilon\mathbb{Z} \tag{3.24}$$

and are known as the two-scale equations. Since $\phi(t)$ in (3.23) and $\psi(t)$ in (3.24) is expressed as linear combination of its shifted half scale versions, $h(k)$ and $g(k)$ defines the weight of each half scale component and are known as the scaling filter and wavelet filter respectively.

3.5. Implementation of Discrete Wavelet Transform

3.5.1 Implementation of Discrete Wavelet Transform using Filterbanks

Recalling the nesting property given in (3.14). One can find that if $\phi(t)$ is in V_0 , then it is also in V_1, the space spanned by $\phi(2t)$. This means $\phi(t)$ can be expressed in terms of a weighted sum shifted $\phi(2t)$ as

$$\phi(t) = \sum_n h(n)\sqrt{2}\,\phi(2t - n), n\epsilon Z \tag{3.25}$$

where, the coefficients $h(n)$ are a sequence of real or perhaps complex numbers called the scaling function coefficients (or the scaling filter or scaling vector) and $\sqrt{2}$ maintains the norm of the scaling function with the scale of two. Expression (3.25) is called the refinement equation, the multi-resolution analysis (MRA) equation or the dilation equation or scaling equation.

Furthermore, according to (3.17), one can find that these wavelets reside in the space spanned by the next narrower scaling function, i.e., $W_0 \subset V_1$. Therefore, they can be also represented by a weighted sum of shifted scaling function $\phi(2t)$ defined in (3.21) by

$$\psi(t) = \sum_n g(n)\sqrt{2}\,\phi(2t-n), n\epsilon Z \tag{3.26}$$

for some set of coefficients $g(n)$. From the requirement that the wavelets span the orthogonal complement spaces as well as the orthogonality of integer translates of the wavelet and the scaling function, the coefficients $h(n)$ and $g(n)$ have the following relations

$$g(n) = (-1)^n h(1-n) \tag{3.27}$$

This subsection will show that the wavelet transform can be implemented via filterbanks that embedded $h(n)$ and $g(n)$ given in (3.25) and (3.26).

3.5.2 Analysis Filterbanks for Forward Wavelet Transform

The discrete wavelet transform can be implemented using the filterbank structure as described below. Replacing t by $\left(2^{-j}t - k\right)$ in (3.25) gives

$$\phi\left(2^j t - k\right) = \sum_n h(n)\sqrt{2}\,\phi\left(2\left(2^j t - k\right) - n\right)$$

$$= \sum_n h(n)\sqrt{2}\,\phi\left(2^{j+1}t - 2k - n\right) \tag{3.28}$$

with $m = 2k + n$, (3.28) becomes

$$\phi\left(2^j t - k\right) = \sum_n h(m-2k)\sqrt{2}\,\phi\left(2^{j+1}t - m\right),\ m, n\epsilon\mathbb{Z} \tag{3.29}$$

$$a_j(k) = \langle s(t), \phi_{j,k}(t)\rangle$$

$$= \int s(t)\, 2^{j/2}\phi\left(2^j t - k\right) dt$$

$$= \int s(t)\, 2^{j/2} \sum_m h(m-2k)\sqrt{2}\phi\left(2^{j+1}t - m\right)\ dt$$

$$= \sum_m h(m-2k) \int s(t)\, 2^{(j+1)/2}\phi\left(2^{j+1}t - m\right) dt$$

$$a_j(k) = \sum_m h(m-2k)\, a_{j+1}(m) \tag{3.30}$$

Expression (3.30) relates the DWT approximation coefficients at the j^{th} and $(j+1)^{\mathrm{th}}$ scales. Similarly, one can obtain the expression of DWT detail coefficients $d_j(k)$ in terms of $a_{j+1}(m)$ as

$$d_j(k) = \sum_m g(m-2k)\, a_{j+1}(m) \tag{3.31}$$

Therefore, from expressions (3.30) and (3.31), one can perform the wavelet transform that does not require explicit forms of $\psi(t)$ and $\phi(t)$ but only depends on $h(n)$ and $g(n)$. In other words, the wavelet transforms can be implemented through the two-channel filterbanks where filtering of a signal by a low pass filter $h(n)$ and high pass filter $g(n)$. Then, the low pass and high pass filter outputs are down sampled by two, respectively; which removes the odd-numbered components after filtering. Consequently, the signal length of the low pass or high pass filter output is only half of original input. This processing is also called analysis bank. The implementation of equations (3.30) and (3.31) is illustrated in Fig. 3.2, where the down-pointing arrows denote decimation by two and boxes $h(-n)$ and $g(-n)$ denotes the finite impulse response (FIR) filters. The FIR filter implemented by $h(-n)$ is a low pass filter (LPF), and $g(-n)$ is a high pass filter (HPF). This algorithm for implementing DWT is known as Mallat's algorithm or pyramid algorithm [57, 58].

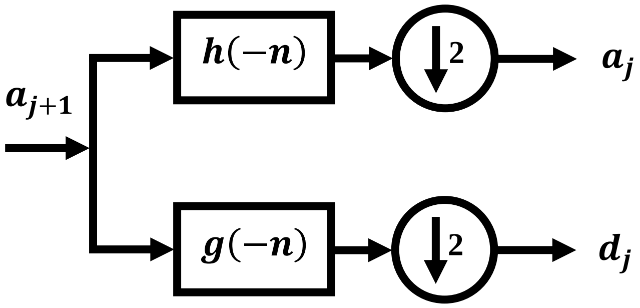


Fig. 3.2: Implementation of DWT by Mallat's algorithm.

The frequency response of a digital filter is the discrete Fourier transform (DFT) of its impulse response $h(n)$, and $g(n)$. This is defined as

$H(\omega) = \sum_{k=-\infty}^{\infty} h(n)e^{j\omega n}$ (3.24)

$G(\omega) = \sum_{k=-\infty}^{\infty} g(n)e^{j\omega n}$ (3.25)

Using this defination, the filter coefficients $h(-n)$ and $g(-n)$ can be re-drawn as $H(\omega)$ and $G(\omega)$ respectively.Using this notation Fig. 3.2 can be re-drawn as

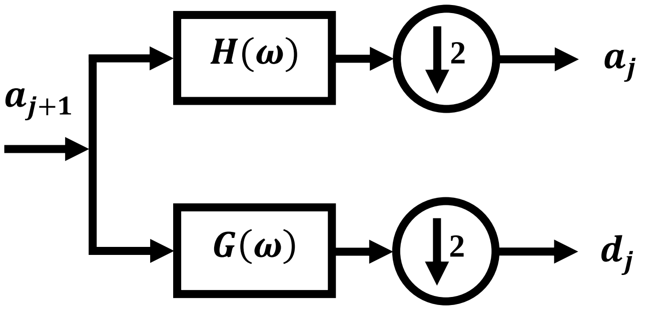


Fig. 3.3: Implementation of DWT by Mallat's algorithm re-drawn.

For the level-3 decomposition, the iterating filterbank structure is shown in Fig. 3.4. The first stage of three banks divides the spectrum of $a_3(k)$ into low pass and high pass bands resulting in the scaling and wavelet coefficients at lower scale $a_2(k)$ and $d_2(k)$. The second stage then divides the low pass band into further lower low pass and band pass bands. This results in a logarithmic progression of bandwidths as illustrated in Fig. 3.3.

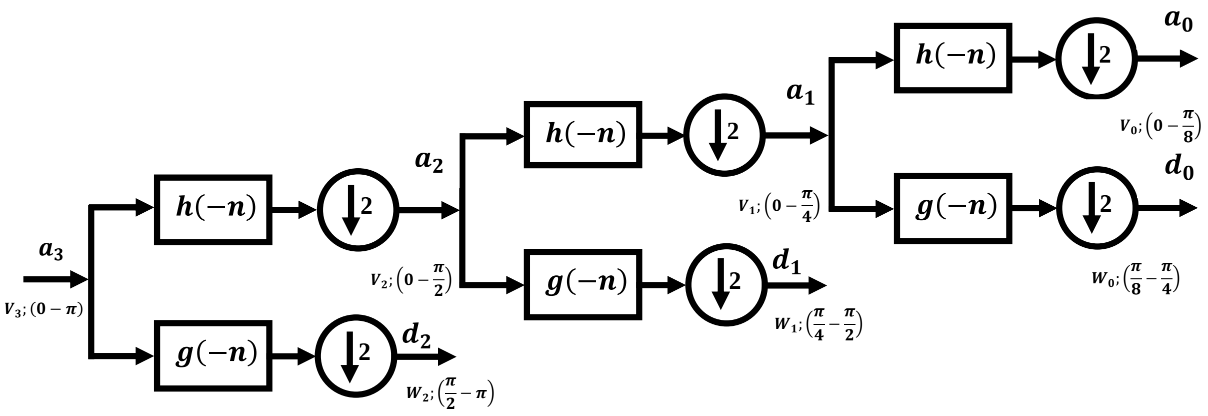

a_3
$V_3; (0-\pi)$
$h(-n)$
$g(-n)$
2
2
a_2
$V_2; \left(0-\frac{\pi}{2}\right)$
d_2
$W_2; \left(\frac{\pi}{2}-\pi\right)$
$h(-n)$
$g(-n)$
2
2
a_1
$V_1; \left(0-\frac{\pi}{4}\right)$
d_1
$W_1; \left(\frac{\pi}{4}-\frac{\pi}{2}\right)$
$h(-n)$
$g(-n)$
2
2
a_0
$V_0; \left(0-\frac{\pi}{8}\right)$
d_0
$W_0; \left(\frac{\pi}{8}-\frac{\pi}{4}\right)$

Fig. 3.4: Three-level DWT decomposition tree.

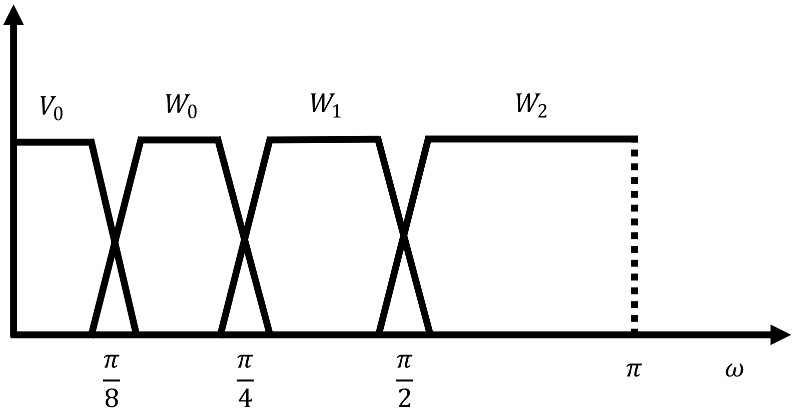


Fig. 3.5: The frequency bands of the three-level DWT decomposition tree.

3.6. Wavelet Packet Transform

In the previous section, we saw that in DWT the approximation (low frequency component) spaces (V_j) were iteratively decomposed until the space V_0 is obtained and whose bias is the scaling function $\phi(t)$. The wavelet packet transform (WPT) is a generalization of the DWT where the detail (high frequency component) spaces (W_j) are also decomposed [59]. From MRA [57], we know that, given the basis functions $\phi_{j_0,k} = \{\phi(t-k)|n \in \mathbb{Z}\}$ and $\psi_{j_0,k} = \{\psi(t-k)|n \in \mathbb{Z}\}$ of V_0 and W_0 respectively, the basis function of V_1 is $\phi_{1,k} = \{\phi(2t-k)|, n \in \mathbb{Z}\}$, where $V_1 = V_0 \oplus W_0$, and $\phi(t)$ and $\psi(t)$ are as defined in (3.23) and (3.24). The splitting algorithm can be used to decompose W spaces as well. For example, if we analogously define $\omega_2(t) = \sqrt{2}\sum_k h(k)\,\psi(2t-k)$, and $\omega_3(t) = \sqrt{2}\sum_k g(k)\,\psi(2t-k)$, then $\{\omega_2(t-k)\}$ and $\{\omega_3(t-k)\}$ are the orthonormal basis function for the two subspaces whose direct sum is W_1.

This can be generalized for $n = 0, 1, 2, 3, \dots, 2^j - 1.$ by defining a sequence of functions, $\{\omega_n(t)\}_{n=0}^{\infty}$ as

$$\omega_{2n}(t) = \sqrt{2}\sum_k h(k)\,\omega_n(2t-k) \tag{3.32}$$

and

$$\omega_{2n+1}(t) = \sqrt{2}\sum_k g(k)\,\omega_n(2t-k) \tag{3.33}$$

where for $n = 0, \omega_0(t) = \phi(t)$ is the scaling function and $\omega_1(t) = \psi(t)$ is the wavelet function [59]. For the case of $J = 2$, the full-tree decomposition is shown in Fig. 3.6. The figure also shows the WPT coefficients and the frequency bands at each node of decomposition. In Fig. 3.7, decomposition of the starting signal space into orthogonal spaces is shown along with the frequency bands for this two-level full-tree WP decomposition.

One of the main ingredients in the wavelets transform is the down-sampling at each scale. Although the down-sampling reduces the output data-rate and results in compact representation, it also introduces one artifact—*shift-variance*. The wavelet transform of a signal and the wavelet transform of a shifted version of the same signal is drastically different. Therefore, the lack of *shift-invariance* is one well-known disadvantage of the discrete wavelet transform and wavelet packet transform [60, 61]. There has recently been much interest in developing shift-invariant orthonormal wavelet packet transform. In Cohen *et* al. [66] the decimation step is adaptively chosen at the current level (even or odd indexed terms are retained).

The stationary wavelet packet transform (SWPT) described in Section 3.7, in practical circular filtering terms, involves no decimation. We maintain full resolution in time and frequency by avoiding the 'best basis' selection which also gives different results according to the somewhat arbitrary choice of cost function.

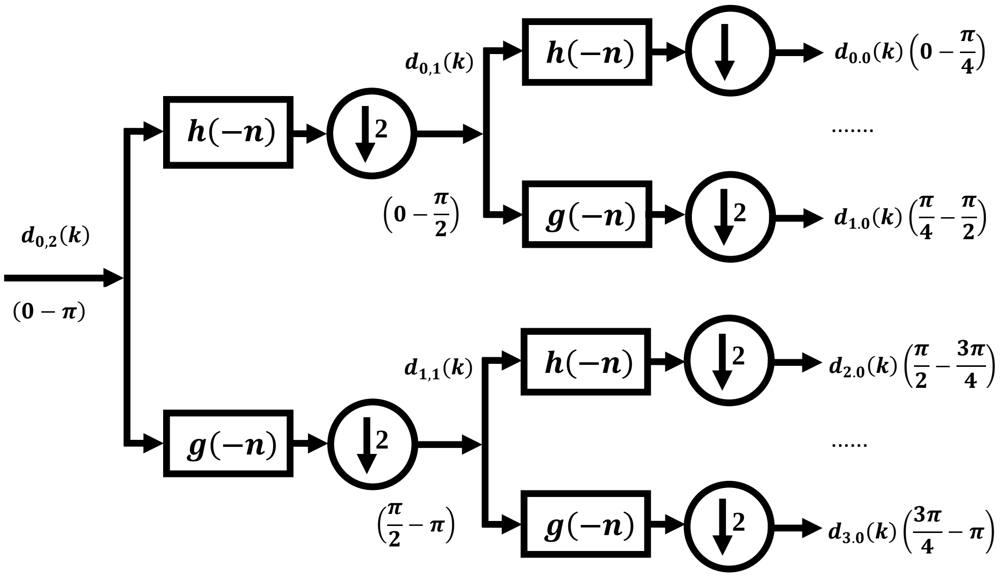

$d_{0,2}(k)$
$(0-\pi)$
$h(-n)$
2
$g(-n)$
2
$d_{0,1}(k)$
$\left(0-\frac{\pi}{2}\right)$
$d_{1,1}(k)$
$\left(\frac{\pi}{2}-\pi\right)$
$h(-n)$
$g(-n)$
2
$h(-n)$
2
$g(-n)$
2
$d_{0.0}(k)\left(0-\frac{\pi}{4}\right)$
....
$d_{1.0}(k)\left(\frac{\pi}{4}-\frac{\pi}{2}\right)$
$d_{2.0}(k)\left(\frac{\pi}{2}-\frac{3\pi}{4}\right)$
......
$d_{3.0}(k)\left(\frac{3\pi}{4}-\pi\right)$

Fig. 3.6: The two-level tree structured filterbank implementation of wavelet packet transform and the frequency bands.

<table>
<tr><td colspan="4">$\boldsymbol{W_{0,2}}$
$\boldsymbol{(0-\pi)}$</td></tr>
<tr><td colspan="2">$\boldsymbol{W_{0,1}}$
$\boldsymbol{\left(0-\frac{\pi}{2}\right)}$</td><td colspan="2">$\boldsymbol{W_{1,1}}$
$\boldsymbol{\left(\frac{\pi}{2}-\pi\right)}$</td></tr>
<tr><td>$\boldsymbol{W_{0,0}}$
$\boldsymbol{\left(0-\frac{\pi}{4}\right)}$</td><td>$\boldsymbol{W_{1,0}}$
$\boldsymbol{\left(\frac{\pi}{4}-\frac{\pi}{2}\right)}$</td><td>$\boldsymbol{W_{2,0}}$
$\boldsymbol{\left(\frac{\pi}{2}-\frac{3\pi}{4}\right)}$</td><td>$\boldsymbol{W_{3,0}}$
$\boldsymbol{\left(\frac{3\pi}{4}-\pi\right)}$</td></tr>
</table>

Fig. 3.7: The orthogonal spaces and frequency bands of the two-level tree structured wavelet packet transform.

3.7. Stationary Wavelet Packet Transform

The stationary wavelet transform (SWT) and stationary wavelet packet transform (SWPT) is designed to overcome the shift/translation-invariant problem by removing the down-sampling at each decomposition level [62-65]. Thus, the approximation coefficients and detail coefficients at each level are the same length as the original signal. The SWPT is also known as un-decimated wavelet packet transform (UDWPT) or maximal overlap wavelet packet transforms (MOWPT) or shift/translation invariant wavelet packet transform.

The SWPT decomposes a signal into a low frequency subband and a high frequency subband by using two channels filterbank without employing decimation after filtering. Then, the low frequency subband as well as the high frequency subband can be decomposed into a second level, low and high frequency subband and the process is repeated as in Fig. 3.8. At each level, the filter is up-sampled versions of the previous ones. The absence of a decimation leads to a full rate decomposition. Each subband contains the same number of samples as the input. So, for a decomposition of j levels, there is a redundant ratio of $(2^j: 1)$. However, the absence of a decimation makes the SWPT shift-invariant and linear. The SWPT not only improves the frequency resolution, but also maintains a temporal resolution.

For any signal $s(t)$, the discrete stationary wavelet packet coefficient at level j and subband k, with the number of $2^j - 1 + k$, is

$$w_{j,k}(t) = \sum_{\tau=0}^{L_j-1} f_{j,k}(\tau) s[(t-\tau) \bmod N] \tag{3.34}$$

$$d_{j,k}(t) = \sum_{\tau=0}^{L_j-1} f_{j,k}(\tau) w_{j,k}[(t-\tau) \bmod N] \tag{3.35}$$

where $\{f_{j,k}(\tau)\}$ is the stationary wavelet packet filter at level j and subband k [60, 61].

The SWPT is based on filters H and G and on an up-sampling operator. The filter H is a low pass filter defined by a sequence $h(n)$ and the high pass filter G defined by a sequence $g(n)$. The structure of two level SWPT is shown in Fig. 3.8. These filters can be obtained by the inner product of the scaling function $\psi(t)$ and wavelet function $\phi(t)$ as:

$$H(k) = \langle 2^{-1}\psi(2^{-1}t), \psi(t-k)\rangle \tag{3.36}$$

$$G(k) = \langle 2^{-1}\phi(2^{-1}t), \psi(t-k)\rangle \tag{3.37}$$

where $t \epsilon \mathbb{L}^2(\mathbb{R})$ and $k \epsilon \mathbb{Z}$.

The SWPT coefficient frequency ranges for two levels decomposition are shown in Fig. 3.9. At every level, the SWPT frequency resolution is $f_r = \frac{f_s}{2^{j+1}}$ and the frequency bandwidth of SWPT coefficients is $\left[\frac{nf_s}{2^{j+1}}, \frac{(n+1)f_s}{2^{j+1}}\right]$, where $n, \left(n = 0,1,\dots,2^j - 1\right)$ denotes the frequency band index within level j [57, 66-68].

The SWPT or UDWPT is defined in terms of an un-decimated filterbank implementation which is a generalization of Mallat's multi-resolution algorithm for computing DWT [25]. The key point is that we don't down sample the filtering output and keep separately filtering even samples and odd samples from every band.

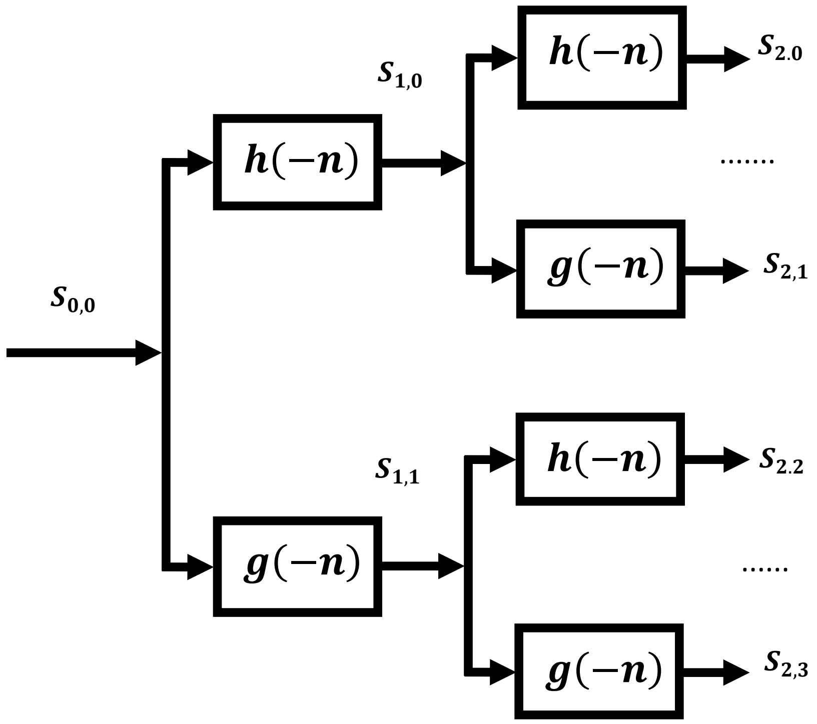


Fig. 3.8: Two level stationary WPT decomposition.

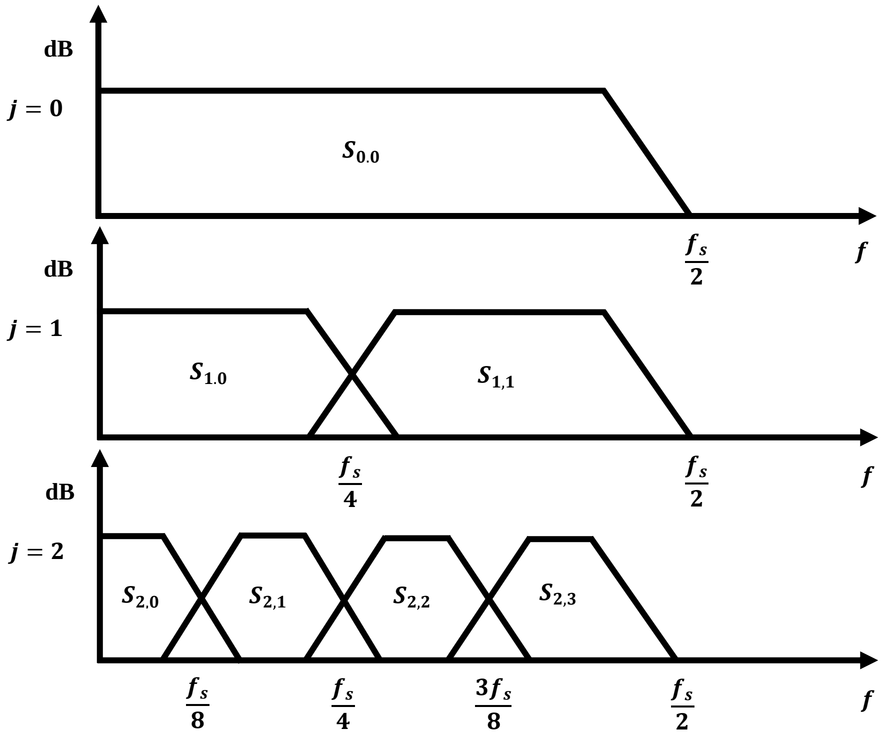


Fig. 3.9: Stationary WPT coefficients frequency range.

3.8. Summary

In this chapter, a detailed study on wavelet transform is presented. Due to efficient time-frequency localization and multi-resolution analysis, the wavelet transform, wavelet packet transform, and stationary wavelet packet transform are suitable for processing the non-stationary signals such as speech [66, 67]. Based on the wavelet framework described in this Chapter, a perceptually motivated stationary wavelet packet transform (WPT) is constructed which forms the basis of the speech enhancement algorithm proposed in Chapter 6 of this thesis.

The next chapter describes an iterative processing based multi-band spectral subtraction speech enhancement algorithm which is based on multi-band spectral subtraction algorithm as described in Chapter 2.

Chapter 4

Iterative Processing based Multi-Band Spectral Subtraction

4.1. Introduction

In real-world environments, the noise signal affects the speech spectrum differently at various frequencies. Therefore, for enhancement of speech degraded by real-world noise (which is usually of non-stationary), a multi-band spectral subtraction algorithm is found to be more appropriate [21]. Although the multi-band approach enhances the overall quality of speech to a reasonable extent, some unnatural sound tones does remain in the enhanced speech. This noise is referred as remnant noise which is of perceptual annoying nature and causes listening fatigue.

To reduce the remnant noise further, in this chapter we propose a novel speech enhancement algorithm that performs iterative processing for the multi-band spectral subtraction (MBSS) algorithm [21]. The iterative processing suppresses the remnant noise further, by iterating the enhanced output signal to the input again and performing the operation, repeatedly.

The work reported in this chapter has resulted in the following publication

[I]. Navneet Upadhyay, and Abhijit Karmakar, "A multi-band spectral subtraction using iterative processing for enhancement of degraded speech," in *Proceedings of International Conference on Signal, Image, and Video Processing*, IIT Patna, India, Jan. 13─15, 2012, pp. 192─196.

[II]. Navneet Upadhyay, and Abhijit Karmakar, "Single channel speech enhancement utilizing iterative processing of multi-band spectral subtraction algorithm," in *Proceedings of 2nd IEEE International Conference on Power, Control, and Embedded System*, MNNIT Allahabad, Dec. 17─19, 2012, pp. 196─201.

[III]. Navneet Upadhyay, and Abhijit Karmakar, "A multi-band speech enhancement algorithm exploiting iterative processing for enhancement of single channel speech," *Journal of Signal and Information Processing*, 2013. [Accepted, Under Print]

The output from the MBSS algorithm that contains remnant musical noise is used as the input signal again for next iteration process. After the first MBSS processing step, the additive background noise transforms to the remnant musical noise. Therefore, the remnant noise needs to be further re-estimated. The newly estimated remnant noise is further used to process the next MBSS step. This procedure is iterated a small number of times. The simulation results as well as the objective and subjective evaluations, explained in Section 4.5, confirm that the enhanced speech obtained by the proposed algorithm is more pleasant to listeners than speech enhanced by conventional MBSS algorithm [21].

The rest of the chapter is organized as follows. In Section 4.2, the proposed algorithm, an iterative processing based multi-band spectral subtraction, is described. In Section 4.3, the details of the performance measures, both subjective and objective, are elaborated. In Section 4.4, the details about speech and noise database are given which is used for evaluation of proposed algorithm. Section 4.5 presents the study of speech enhancement results and finally Section 4.6 concludes this chapter.

4.2. Iterative Processing based Multi-Band Spectral Subtraction Algorithm

In order to suppress the remnant noise, produced by the multi-band spectral subtraction algorithm, as explained in Section 2.5.2, in a real-world environment, we have used the multi-band spectral subtraction algorithm [21] that makes use of the proposed iterative processing. The iterative processing is a technique in which the speech enhancement procedure is executed on the estimated speech that is taken as the input and processed repeatedly, to obtain the further enhanced speech and thus reducing the remnant noise. The reduction of remnant noise can be achieved by estimating noise in each iteration, and improving the quality of speech progressively. The iterative processing method can also be closely related to the Wiener filtering based speech enhancement method [22, 38], as explained below.

The block diagram of the proposed iterative processing based multi-band spectral subtraction, IP-MBSS, is illustrated in Fig. 4.1. In the proposed algorithm, the speech enhanced by MBSS algorithm is used as the input of the next iteration process. It is evident from Fig. 4.1 that the additive background noise changes its form to remnant noise after the completion of the initial step of reference MBSS algorithm. For example, say $y(n)$ is the input signal and after the MBSS processing the obtained enhanced speech is say $\hat{s}(n)$, as described in Section 2.5.1.2. Thus, the additive noise is greatly reduced by the MBSS algorithm. This reduction in noise is associated with the presence of irritating remnant noise which is of musical structure in the enhanced speech, $\hat{s}(n)$. In our proposed algorithm, this enhanced speech i.e., output signal is used as the input signal again for the next iteration process by re-estimating the remnant noise from each band in each iteration. Therefore, the final enhanced output speech signal can be obtained after a finite number of iteration steps.

Moreover, if we regard the process of noise estimation and the multi-band spectral subtraction as a filtering step, then the output is used not only for designing the filter but also used as the input signal of the next iteration process. More importantly, this filter can be refreshed adaptively by re-estimating the remnant noise to improve the speech quality and intelligibility, effectively.

Let us assume that the signal at the m^{th} iteration step is given by

$$y(m,n) = s(m,n) + d(m,n), \quad n \epsilon (0, N-1) \tag{4.1}$$

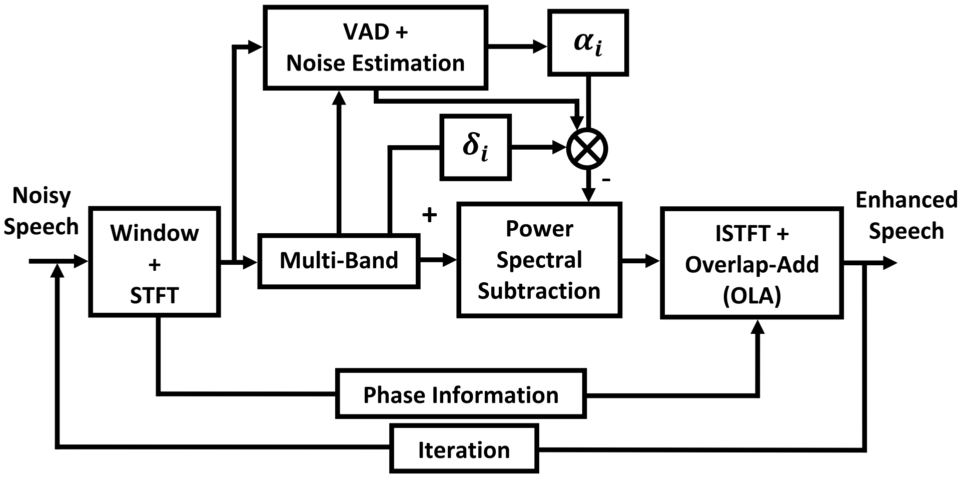


Fig. 4.1: Block diagram of an iterative processing based multi-band spectral subtraction algorithm.

where, ω_i and ω_{i+1} are the start and end frequency bins of the i^{th} Band; α_i is the band specific over-subtraction factor of the i^{th} Band; δ_i is an additional band subtraction factor for each band. The $\left|\hat{S}_i(m,\omega)\right|^2$, $|Y_i(m,\omega)|^2$, and $\left|\widehat{D}_i(m,\omega)\right|^2$ is the power spectrum of estimated speech, noisy speech and estimated remnant noise power in the i^{th} Band at the m^{th} iteration step, respectively. The estimate of the complete speech signal in the m^{th} iteration $\hat{s}(m,n)$ is obtained by performing overlap-add processing (see Appendix A) after obtaining $\left|\hat{S}_i(m,\omega)\right|$ from (4.2). Here, the phase of the noisy speech signal is used to obtain, $\hat{S}_i(m,\omega)$.

Here, we note that (4.2) can be written as

$$\hat{S}_i(m,\omega) = \left|\hat{S}_i(m,\omega)\right|.\frac{Y_i(m,\omega)}{|Y_i(m,\omega)|} \tag{4.3}$$

$$\hat{S}_i(m,\omega) = |G(m,\omega)|.Y_i(m,\omega) \tag{4.4}$$

where

$$|G(m,\omega)| = \sqrt{1 - \frac{\alpha_i.\delta_i.|\widehat{D}_i(m,\omega)|^2}{|Y_i(m,\omega)|^2}} \tag{4.5}$$

where $\left|\hat{S}_i(m,\omega)\right|$, $|Y_i(m,\omega)|$, and $|G(m,\omega)|$ represent the magnitude of the estimated speech, speech with remnant noise and gain at m^{th} iteration step. We note that (4.4) is having a direct correspondence with Wiener filtering equation, as explained in Section 2.5.3. In the $(m+1)^{\text{th}}$ iteration processing step, the output signal $\hat{s}(m,\omega)$ that is obtained by the m^{th} iteration processing step is set as the input signal again as follows:

$$y(m+1,n) = \hat{s}(m,n) \tag{4.6}$$

Here, the noise component of $y(m+1,n)$ transforms into the remnant noise component that could not be suppressed by the MBSS at m^{th} iteration. The IP-MBSS algorithm obtains the estimated noise spectrum $\left|\widehat{D}_i(m+1,\omega)\right|$ that is to be used in each band of $y(m+1,n)$. If M denotes the final number of times the iterations are repeated, then the resultant estimated speech is denoted by $\hat{s}(M,n)$. As the amount of the noise component is reduced in each MBSS processing step, with an increase in the number of iterations will reduce the amount of noise, progressively.

In summary, the steps of the proposed IP-MBSS algorithm are given below:

STEP 1: *Initialize the iteration count* m *to* $m = 1$.

$y(1, n) = y(n)$

STEP 2: *For* $m = 1$, *estimate the noise spectrum* $\left|\widehat{D}_i(m, \omega)\right|$ *from the silent periods of* $y(n)$ *. On the other hand, for* $m \neq 1$, *estimate* $\left|\widehat{D}_i(m, \omega)\right|$ *from the silent periods of* $\hat{s}(m - 1, n)$, *which was obtained according to iterative processing.*

STEP 3: *Compute the band specific over-subtraction factor* α_i , *also set the additional band subtraction factor* δ_i *for each band empirically. Multiply the scaling factor* $\alpha_i . \delta_i$ *by the estimated noise spectrum* $\left|\widehat{D}_i(m, \omega)\right|$ *to adjust the amount of reduction in each band.*

STEP 4: *Use the estimated noise spectrum with the scaling factor* $\alpha_i . \delta_i$ *, which was obtained in* STEP *3, to execute MBSS processing in each band for* $y(n)$, *when* $m = 1$ *or for* $\hat{s}(m - 1, n)$ *when* $m \neq 1$. *In addition, obtain the output signal* $\hat{s}(m, n)$ *in the* m^{th} *iteration.*

STEP 5: *Set* $m = m + 1$ *and repeat the processing in steps.*

STEP 6: *Stop, if remnant noise is within acceptable limit. Otherwise, go to* STEP *5.*

The number of iteration steps is an important parameter of this proposed algorithm which affects its performance. The segmental SNR (SegSNR) at the end of each iteration step depends on the iteration number and the segmental SNR increases with the number of iterations. Thereby, the over-subtraction factor also increases as it depends on SegSNR directly. In Fig. 4.2, the variation of over-subtraction factor (mean value) with iteration number has been shown and explained in Section 4.5.1 of this chapter. The larger iteration number is expected to give better speech enhancement performance containing less remnant noise. But the performance of the algorithm deteriorates after a certain number of iterations. This will be elaborated in section in Section 4.5.1.

4.3. Performance Measures

4.3.1. Subjective Measures

Subjective measures are performed via listening tests. A review can be found in [3, 11, 69-71]. Subjective intelligibility measures lead to intelligibility scores and subjective quality measures to an overall impression. These tests can be informal or follow a determined protocol. They differ by the following points: i) The choice of speech material (sentences, phonemes, words) which can be nonsense or meaningful, ii) the formant of the test, and iii) the scoring method (open set or closed set, scaling). The principal existing subjective measures are the following:

1) **Intelligibility tests :** Diagnostic Rhyme Tests (DRT) and Modified Rhyme Test (MRT). A rhyme test is a closed set response test in which the listener has to select his response among a small group of possible words (2 alternatives for the DRT, 6 alternatives for the MRT). The DRT tests only the initial consonant. The percentage of correctly heard words is a measure of intelligibility. A good score for intelligibility is in the range 85-90%. The DRT is given by the following relation:

   $$\text{DRT} = \frac{N_{\text{correct}} - N_{\text{incorrect}}}{N_{\text{test}}} \times 100\ [\%] \tag{4.7}$$

   where N_{correct} is the number of correctly understood words, $N_{\text{incorrect}}$ is the number of incorrect words and N_{test} is the total number of words used for the test.

2) **Quality tests:** use of sentences in order to obtain an overall impression on intelligibility, acceptability, naturalness, etc. The most often used measures are the following:

   a) Diagnostic Acceptability Measure (DAM). This test is composed of 12 quality measures on a scale 0-100.

   b) Mean Opinion Score (MOS). This test provides scores on a 5-point scale as described in Table 4.1 [3, 11]. This method does not give absolute values but is well suited for ranking. Sometimes it is necessary to normalize the differences between MOS evaluations.

The MOS does not give any idea of the type of distortion, but only of the degree of impairment. Therefore, very different noise types can lead to the same quality. The DAM gives more complete results but is more time consuming.

TABLE 4.1. MEAN OPINION SCORE FIVE-POINT SCALE

Rating	**Speech Quality**	**Level of Distortion**
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but objectionable
1	Bad/Un-satisfactory	Very annoying and objectionable

Speech quality tests are usually performed on speech with high intelligibility. When speech quality is 'good' or 'excellent', intelligibility is always acceptable (>90%). Speech Intelligibility degrades when the quality is 'poor' or 'bad'. Therefore, in adverse environments (SNR < 10 dB), intelligibility enhancement becomes important.

Subjective tests provide very good results but they are time and money consuming and also difficult to reproduce in the same conditions. It is, therefore, desirable to develop objective measures based on physical aspects of the speech signal to predict the subjective performances. Such objective measures have been developed to replace listening tests. These measures are described in Section 4.3.2.

4.3.2. Objective Measures

A performance measure has to be consistent with human perception. It is, therefore, important to have objective quality measures that are highly correlated to subjective tests, although this is not the general case. Many different objective measures have been developed, especially in the context of speech coding. This section presents an overview of existing measures. The correlation between these measures and subjective listening tests is discussed in Section 4.3.3.

The choice of a performance measure is system dependent. For example, if the end user is a human, the aim of the system will be to improve perceptual aspects of speech. Therefore, the objective measures used for the evaluation will have to show a high degree of correlation with subjective results. On the other hand, if the end user is a recognition system, the aim will be to reduce the recognition error rate. In the second case, the recognition accuracy becomes the objective performance measure.

Generally, objective measures for determining the subjective quality of a speech processing algorithm perform well only for some types of processing. Furthermore, these measures have to be validated by subjective quality and intelligibility tests. It is important to find objective measures well correlated with subjective results.

An objective measure represents a distance from the original speech parameters in noise free conditions [72]. Each measure calculates a distance d_m between the original speech $s(n)$ and an enhanced speech $\hat{s}(n)$. Measures are calculated for each frame m. The averaging of local distances d_m across the whole sentence (all frames) produces a global measure D:

$$D = \frac{1}{M}\sum_{m=0}^{M-1} d_m \tag{4.8}$$

where M represents the number of frames in a signal. Most of the work for calculating d_m has been done on spectral distance measures but there exist a great variety of distances. Distances can be linear or logarithmic, and frequency weighing can improve the correlation with subjective results by giving higher weights to perceptual important components. Here is a description of some of objective measures:

1) **Global signal-to-noise ratio:** one of the oldest and widely used objective measures, defined as the ratio of the total signal energy to the total noise energy in the utterance. It is mathematically simple to calculate, but requires both distorted (noisy) and undistorted (clean) speech samples.

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_{n=1}^{L} s^2(n)}{\sum_{n=1}^{L}\{s(n)-\hat{s}(n)\}^2}\right) \tag{4.9}$$

where $s(n)$ is the clean speech signal, $\hat{s}(n)$ is the enhanced speech reproduced by a speech processing system, n is the sample index, and L is the number of samples in both speech signals. The summation is performed over the signal length.

But, the noise has different impacts on the various sections of speech signal. Therefore, the enhancement criterion as well as the performance evaluation depends on the impact of noise. In this context, a global SNR measure doesn't reflect the impact of noise on each specific section. This classical definition of the SNR does not correlate well with speech; as the SNR averages the ratio over the entire signal. This objective quality measure has a very poor correlation with subjective results.

2) **Segmental signal-to-noise ratio:** this measure makes a conversion of the SNR into dB prior to averaging it, in order to give equal weights to loud and soft part of speech. Speech energy fluctuates over time, and so portions where speech energy is large, and noise is relatively inaudible, should not be washed out by other portions where speech energy is small and noise can be heard over speech. Thus, several variations to the classical SNR exist which show much higher correlation with subjective quality. Therefore, the SNR is calculated in short frames, and then averaged. This measure is called the segmental SNR (SegSNR), and is defined as

$$\mathrm{SegSNR} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\left(\frac{\sum_{n=N_m}^{N_m+N-1} s^2(n)}{\sum_{n=N_m}^{N_m+N-1}\{s(n)-\hat{s}(n)\}^2}\right) \tag{4.10}$$

where M represents the number of frames in a signal and N the number of samples per frame. In order to improve the measure, it is possible to suppress the periods of silence before calculating the SNR. The frame based SegSNR equation is reasonable measure of speech quality or takes into account both remnant noise and speech distortion. Since the logarithm of the ratio is calculated before averaging, the frames with an exceptionally large ratio is somewhat weighed less, while frames with low-ratio is weighed somewhat higher. It can be observed that this matches the perceptual quality well, i.e., frames with large speech and no audible noise does not dominate the overall perceptual quality, but the existence of noisy frames stands out and will drive the overall quality lower. However, if the speech sample contains excessive silence, the overall SegSNR values will decrease significantly since silent frames usually show large negative SegSNR values. In this case, silent portions should be excluded from the averaging using speech activity detectors. In the same manner, exclusion of frames with excessively large or small values from averaging generally results in SegSNR values that agree well with the subjective quality. Furthermore, there exists a frequency weighted segmental SNR [3, 11].

3) **SNR improvement or gain:** SNR improvement indicates the difference between the SNR of enhanced speech (output) and the noisy speech signal (input) or segmental SNR

$$\text{SNR improvement} = \text{SNR}\,\{\hat{s}(n)\} - \text{SNR}\,\{s(n)\}\ \ [\text{dB}] \tag{4.11}$$

$$\text{SegSNR improvement} = \text{SegSNR}\,\{\hat{s}(n)\} - \text{SegSNR}\,\{s(n)\}\quad [\text{dB}] \tag{4.12}$$

In above equations, the first term corresponds to output SNR and second term corresponds to input SNR. For the calculation of input SNR, the denominator term $\{s(n) - \hat{s}(n)\}^2$ of (4.9) is replaced by $\{d(n)\}^2$.These equation takes into account both remnant noise and speech distortion.

4) **Itakura–Saito Distortion Measure:** Itakura-Saito indicates the perceptual difference between an original spectrum $S(\omega)$ and an approximation $\hat{S}(\omega)$ of that spectrum. The ISD is defined as the distance between two log-scaled DFT spectra averaged over all frequency bins. The distance is defined as

$$\text{ISD}\left(S(\omega), \hat{S}(\omega)\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left\{\frac{S(\omega)}{\hat{S}(\omega)} - \log\left(\frac{S(\omega)}{\hat{S}(\omega)}\right) - 1\right\} d\omega \tag{4.13}$$

A typical range for the ISD measure is 0~10, where the minimal value of ISD corresponds to the best speech quality. A correlation between the ISD and subjective quality measures is given in [3, 73, 87].

5) **Perceptual Evaluation of Speech Quality:** Perceptual Evaluation of Speech Quality (PESQ—an objective way of measuring the subjective speech quality) is an objective quality measure algorithm designed to predict the subjective opinion score of a degraded audio sample. It is recommended by ITU-T for speech quality assessment [74, 75]. In PESQ measure, a reference signal and the processed signal are first aligned in both time and level. This is followed by a range of perceptually significant transforms which

include Bark spectral analysis, frequency equalization, gain variation equalization and loudness mapping. After the two signals have undergone these transformations, two parameters are computed. These parameters are then combined in a mapping function to give an estimate of mean opinion score. For normal subjective test material the PESQ score ranges from 0 to $+4.5$, although for most cases it will be a MOS-like score between 1 and 4.5, with higher score indicating better quality. The PESQ measure was reported to be highly correlated with subjective listening tests in [74, 75] for a large number of testing conditions.

4.3.3. Correlation with Subjective Results

Subjective listening tests are the most reliable performance measure, as the end user is a human. However, when developing a new algorithm in various noise conditions, subjective tests become difficult to realize. In this case it would be desirable to replace them by objective measures. Furthermore, objective measures allow an identification of the most distorted speech segments, and this information is very useful for improving speech enhancement algorithms [3].

In order to replace subjective listening tests, the objective measures have to show a high degree of correlation with subjective results. This correlation depends on the type of distortion such as in the case of narrowband or wideband speech. A correlation of $\rho = 1$ means a perfect prediction of the subjective results. Table 4.2 represents a typical case of the correlations obtained for different objective quality measures [3, 70]. The first two objective measures in the table (SNR and segmental SNR) have been obtained for distortions evaluated in waveforms coders. The last objective measure in the table, i.e., PESQ shows maximum correlation with the subjective listening tests or equivalently, MOS scores. Extreme care must be taken when comparing these values, as they result from different test conditions and different distortions. In general, log measures have a higher correlation than the corresponding linear measures. Most of these measures have free parameters that can be optimized to maximize the correlation ρ. This optimization is usually performed during a training phase with speech samples of known subjective quality. In [3], it is stated that a correlation of $\rho = 0.97$ or more is necessary for a reliable measure. This is achieved only for measure that takes into account facts about human perception. However, none of the existing measures is able to replace subjective tests in all conditions, but they can be used to obtain preliminary results in a development phase. This shows the difficulty of defining a measure able to predict the subjective quality for each type of distortion and each type of enhancement algorithm.

TABLE 4.2: CORRELATION COEFFICIENTS BETWEEN SUBJECTIVE AND OBJECTIVE MEASURES

Objective measures	Correlation coefficients
SNR (Global) Segmental SNR	$\rho = .24$ $\rho = .77$
Itakura-Saito	$\rho = .59$
PESQ	$\rho = .97$

4.4. Speech and Noise Databases

This section describes the database used for the evaluation of the proposed enhancement algorithm, IP-MBSS, for obtaining the performance results which are given in Section 4.5. The speech sentences and noisy speech samples have been taken from NOIZEUS corpus speech database [76]. The NOIZEUS corpus is a publicly available speech database, often used for benchmark experiments. The NOIZEUS corpus is composed of 30 phonetically balanced sentences belonging to six speakers, three male and three female. The speech data are sampled at 8 kHz and quantized linearly using 16 bits resolution.

The noise signals have different time-frequency distributions, and have different impact on speech samples. The sentences in the database are degraded by seven different non-stationary noises at varying SNR levels i.e., at global SNR levels of 0 dB to 5 dB. The non-stationary noises such as car noise, train noise, restaurant noise, babble noise, airport noise, street noise, and exhibition noise are the various noises in the NOIZEUS.

For the evaluation of proposed speech enhancement algorithm, a total of four different sentences pronounced by three male and a female speaker have been taken from NOIZEUS corpus. Every sentence has a silence segment in the beginning that lasts more than 0.25 ms, approximately. Eight different types of noises, seven real-world noise (non-stationary) and a computer generated white Gaussian noise (stationary), have been used for the evaluation and comparison of speech enhancement algorithm. Therefore, the sentences are degraded with eight types of noises at varying SNR levels i.e., at global SNR levels of 0 dB to 5 dB. The performance of the proposed speech enhancement algorithm is tested on such noisy speech samples.

4.5. Study of Enhancement Results

4.5.1. General Considerations

This section presents an evaluation of the performance of the proposed speech enhancement algorithm, as well as comparison with other spectral subtractive-type algorithms. This evaluation is performed for the four speech sentences and the eight different noise (non-stationary and stationary) types presented in Section 4.4. The sampling frequency is 8 kHz and the other parameters that have been used for the implementation of the proposed algorithm are as follows:

1) Frame size = 256 samples (32 ms) with 50% overlap (i.e. 128 samples (16 ms)).
2) Hamming window with 256 samples for input signal weighting.
3) FFT length = 256 points FFT.
4) Noise estimation via exponential averaging with $\lambda_D = 0.9$. According to (2.11), this leads to an averaging of 20 frames (320 ms).
5) No. of continuous non-overlapped uniformly spaced frequency bands $K = 4$
   $\{60\text{ Hz} \sim 1000\text{ kHz (Band 1)}, 1\text{ kHz} \sim 2\text{ kHz (Band 2)}, 2\text{ kHz} \sim 3\text{ kHz (Band 3)},$
   $3\text{ kHz} \sim 4\text{ kHz (Band 4)}\}$.

The proposed algorithm has been compared with the multi-band spectral subtraction algorithm (MBSS) to evaluate the performance benefit the iterative processing gives to the enhancement process. MBSS algorithm offers an adaptation of α based on the segmental SNR. The value of over-subtraction factor α is determined using Fig. 2.4 and (2.14), and the value of additional over-subtraction factor δ_i for each band is set as per (2.15). The value of spectral flooring parameter β is taken as 0.003 [21]. In case of IP-MBSS, the iteration time is an important factor, which directly affects the performance of speech enhancement. In order to explore the relationship between the performance of speech enhancement and the iteration times, the variation of the mean over-subtraction factor α of the speech degraded by car noise with iteration number are shown in Fig. 4.2. It can be seen from figure that the value of α increases as the iteration number increases, which suggest the larger iteration number will correspond to the better speech enhancement performance resulting in less remnant noise. However, both the waveforms and the corresponding spectrogram suggest that the larger iteration number would start eliminating some component of the normal speech while reducing the remnant noise effectively. Therefore, the proposed iteration number for the car speech is set to 2 to 3 and the value of other parameters have been taken as same as the reference algorithm.

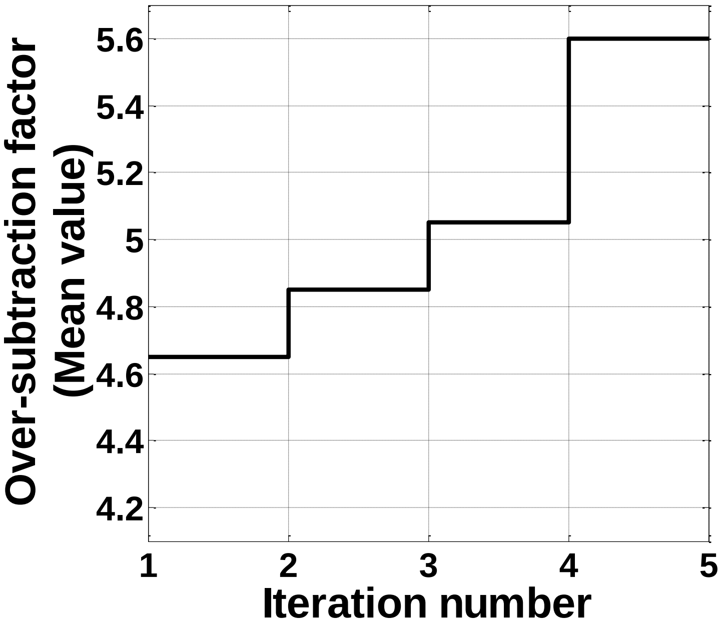


Fig. 4.2: Variations of over-subtraction factor (mean value) with iteration number.

Generally, the objective performance evaluation is based on the application of the objective quality or intelligibility measures described in Section 4.3.2. The major drawbacks of these objective measures are the following: i) they are not always well-correlated with speech perception [3, 73], and ii) they do not give information about how speech and noise are distributed across frequency. Usually, single channel spectral subtractive-type enhancement algorithms produce two main undesirable effects, i.e., remnant musical noise and speech distortion. These effects can be annoying to a human listener, but they are difficult to quantify with the help of these objective measures. It is, therefore, important to analyze the time-frequency distribution of the enhanced speech, in particular, the structure of its remnant musical noise. This is done by observing the speech spectrogram, which gives more accurate information about remnant noise and speech distortion in comparison to the time waveforms.

Taking into account these considerations, the performance evaluation is composed of the following steps:

1) Informal listening tests during the development phase.
2) Observation of time waveforms and speech spectrogram.
3) Measure of the amount of noise reduction.
4) Application of objective intelligibility and quality measures.
5) Analysis of the structure of remnant noise and speech distortion.
6) Subjective listening tests at the end of evaluation in order to validate the results given by the objective measures.

The Section 4.5.2 presents the objective evaluation results of the algorithm, while the next Section 4.5.3 presents the results of subjective listening tests and spectrogram analysis.

4.5.2. Overall Performance in Additive Noise

The additive background noises taken for the tests are car noise, train noise, restaurant noise, babble, airport noise, street noise, exhibition noise and white Gaussian noise with SNR levels ranging from 0 to 5 dB. The measures that have been chosen for the performance evaluation are as follows: signal-to-noise ratio (SNR), segmental signal-to-noise ratio (SegSNR), Itakura-Saito distortion (ISD) and perception evaluation of speech quality (PESQ)

The output SegSNR obtained for various types of noises at various noise levels is presented in Fig. 4.3. The overall trend suggests that the value of output SegSNR for various noise types as mentioned above increases at low input SNRs (≤ 5 dB). It is observed that for the case of exhibition noise our algorithm performs poorly in comparison to other noises at 5 dB SNR, whereas for babble and airport noise it shows very small decrement of SNR. For the case of restaurant noise the output SegSNR gives comparable results with the input.

In Fig. 4.4, the SegSNR improvements for various noise types at various noise levels are presented and compared with MBSS algorithm. The SegSNR improvements provided by the proposed algorithm produces an SNR improvement at low input SNRs (≤ 5 dB) compared to MBSS algorithm while for the case of restaurant and exhibition noises our algorithm shows no improvement in comparison to MBSS algorithm.

Table 4.3 presents the objective evaluation and comparison of the proposed algorithm, IP-MBSS, in terms of output SNR (dB), output SegSNR (dB), and ISD at different labels of SNR. The values of output SNR, output SegSNR for different types of noises of IP-MBSS are observed to be better than MBSS algorithm. In case of ISD, the performance improvement is found to be more for IP-MBSS in compared to MBSS at some places only.

Table 4.4 presents the results of the objective evaluation and the comparison with the proposed algorithm, IP-MBSS, in terms of SNR improvement (dB), and Seg.SNR improvement (dB) at different labels of SNR. The value of SNR improvement, Seg.SNR improvement for different types of noises for IP-MBSS is found to be better than the MBSS algorithm.

4.5.3. Subjective Evaluation and Spectrogram Analysis

In our subjective evaluation, the listening tests have been accomplished with five listeners in a closed room and headphones have been used during experiments. Each listener provides a score between one and five for each test signal. This score represents the listener's overall appreciation of the quality of the speech sample which contains the remnant noise, the left-over background noise and speech distortion. The scale used for these tests corresponds to the MOS scale presented in Table 4.2 [3]. For each speaker, the following procedure has been applied:

1) Clean speech and noisy speech are played and repeated twice;
2) Each test signal, which is repeated twice for each score, is played three times in a random order.

This leads to 20 scores for each test signal and is presented in Table 4.5. As the listeners appreciate speech quality differently, the mean of the different scores varies greatly from one speaker to another. However, the values obtained are well-suited for ranking the performance of the different methods tested. It can be seen from Table 4.5 that the MOS score of the enhanced speech obtained from using the IP-MBSS algorithm is the highest, followed by that from the MBSS algorithm for various sentences taken form NOIZEUS database. Fig. 4.6 shows the spectrogram of the enhanced speech obtained with the proposed algorithm for the speech sentence (sp1) degraded by car noise, train noise, babble noise, restaurant noise, airport, street, exhibition, and white noise, respectively at 5 dB SNR.

Fig. 4.7 shows the temporal waveforms of the enhanced speech obtained with the proposed algorithm for the speech sentence (sp1) degraded by car noise, train noise, babble noise, restaurant noise, airport, street, exhibition, and white noise, respectively at 5 dB SNR. Fig. 4.8 to Fig. 4.12, show the temporal waveforms and spectrogram of the enhanced speech obtained with the proposed algorithm for the speech sentence pronounced by male speaker (sp1, sp6, sp10), female speaker (sp12) and degraded by car noise at 5 dB SNR and 10 dB SNR in comparison with MBSS and BSS algorithms along with PESQ scores. The temporal waveforms and spectrogram of Fig. 4.12 presents the spectrogram of the enhanced speech obtained by IP-MBSS for the sentence pronounced by a female speaker. This results are comparable to the one obtained for a male in similar noise conditions in Fig. 4.8 to Fig. 4.11 along with PESQ scores.

It can be seen from Fig. 4.5 to Fig. 4.12 that the musical structure of the remnant noise is reduced more by the proposed algorithm compared to MBSS and BSS algorithms. Speech enhanced with the proposed algorithm is more pleasant and the remnant noise has a *"perceptually white quality*" while distortion remains within acceptable limit. This is confirmed from the values obtained from the objective measures (Table 4.3 and Table 4.4) and also validated by subjective listening tests.

The results shown in Table 4.5, presents the MOS Scores and PESQ scores of IP-MBSS and MBSS algorithms. It’s clearly evident that, in comparison with MBSS algorithm, the quality of subjective rating of the enhanced speech by the proposed algorithm is much better. In the case of the PESQ measure, the proposed IP-MBSS algorithm gives better PESQ scores than the MBSS and BSS algorithm, as expected.

In Fig. 4.5, a scatter plot is shown between the MOS and PESQ scores of proposed algorithm, IP-MBSS, for various types of real-world noises and a computer generated white Gaussian noise. It is observed that the subjects tend to give higher scores than the PESQ scores, although a high correlation between objective and subjective scores are noticed. It is evident from the figure that the PESQ score and the MOS score correlate poorly for of train and babble noises in all conditions, compared to other types of noises.

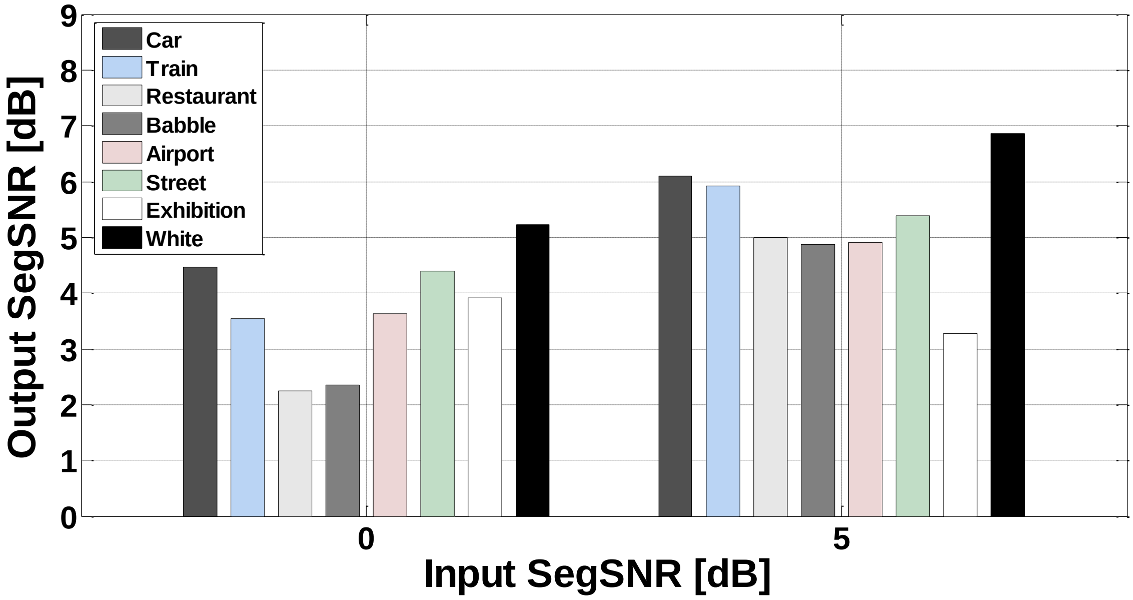


Fig. 4.3: Output SegSNR of IP-MBSS for car, train, restaurant, babble, airport, street, exhibition, and white Gaussian noises.

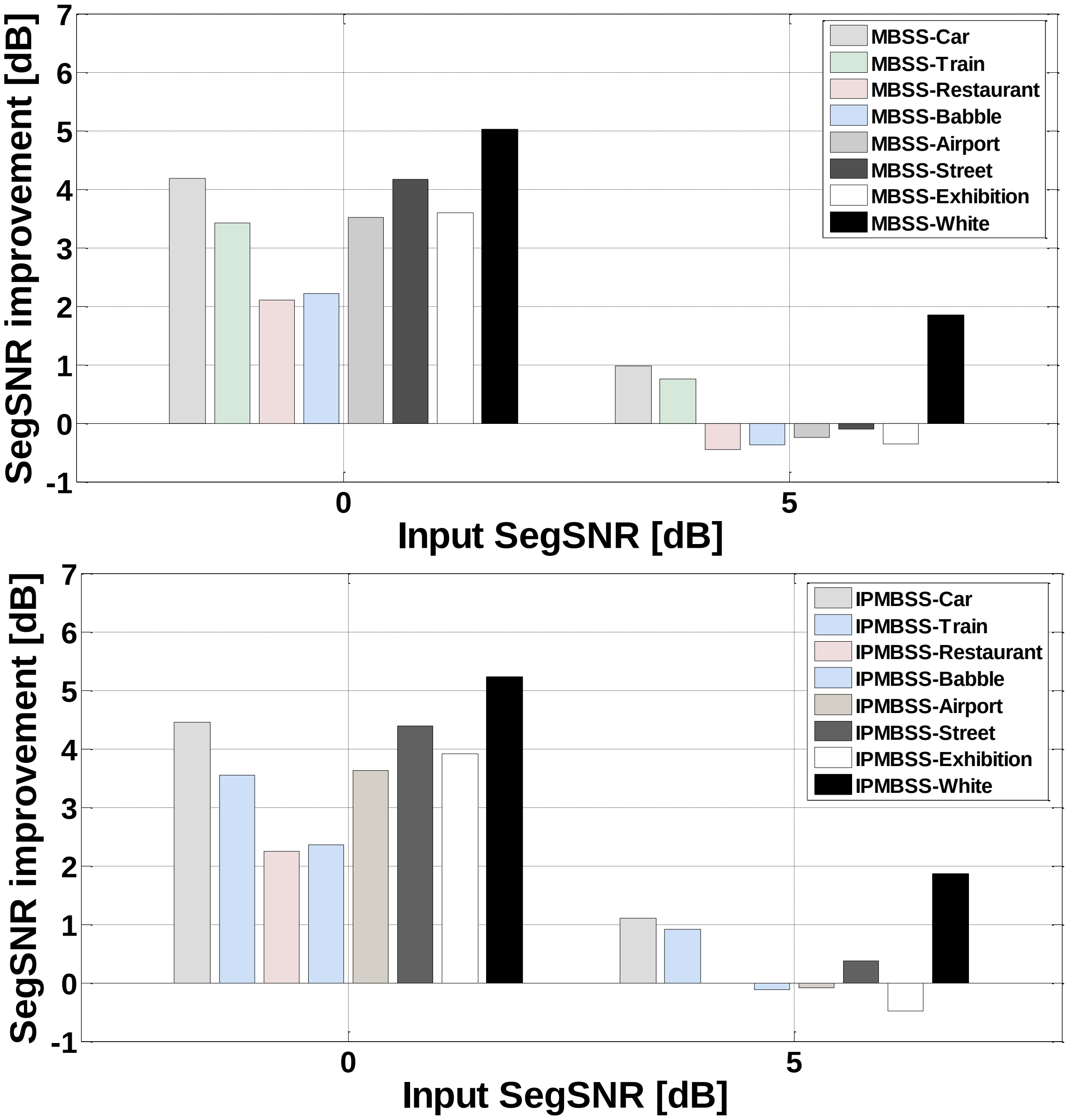


Fig.4.4: SegSNR improvement of IP-MBSS over MBSS for various noise types e.g. car, train, restaurant, babble, airport, street, exhibition, and white Gaussian noises.

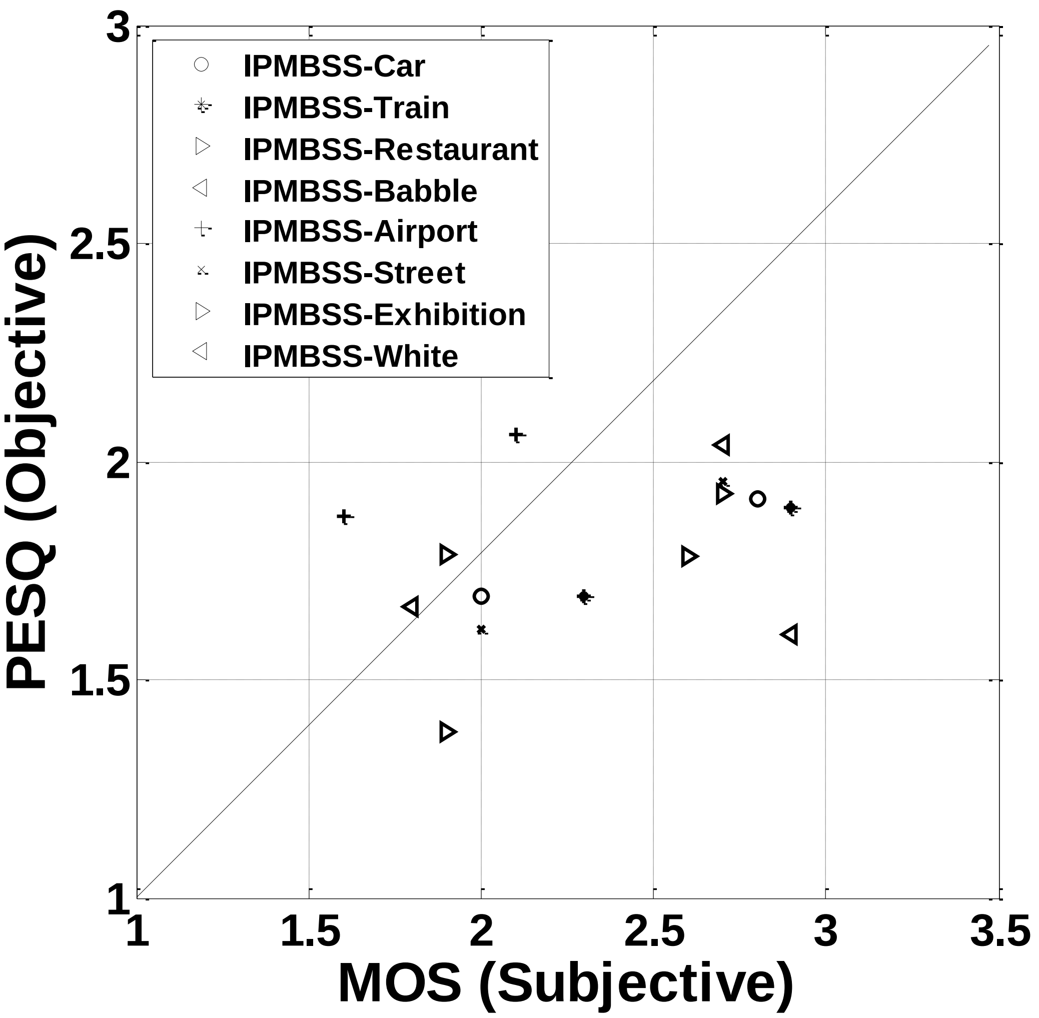


Fig. 4.5: Scattered plot of PESQ score vs. mean MOS score of IP-MBSS for car, babble, restaurant, train, airport, street, exhibition, and white noises.

.

Table 4.3.

Objective Evaluation and Comparison of the Proposed Algorithm in Terms of Output SNR (dB), Output SegSNR (dB), and ISD.

Noise Type	Enhancement Algorithms	SNR (dB)		SegSNR (dB)		ISD	
		0dB	5dB	0dB	5dB	0dB	5dB
Car	MBSS	4.26	6.01	4.19	5.98	1.80	1.40
	IP-MBSS	**4.50**	**6.11**	**4.46**	**6.10**	**1.98**	**1.79**
Train	MBSS	3.47	5.82	3.42	5.75	2.08	1.32
	IP-MBSS	**3.57**	**5.96**	**3.54**	**5.92**	**1.97**	**1.75**
Restaurant	MBSS	2.15	4.60	2.10	4.54	1.49	1.03
	IP-MBSS	**2.27**	**5.04**	**2.24**	**4.99**	**2.11**	**2.02**
Babble	MBSS	2.27	4.64	2.21	4.63	1.77	1.13
	IP-MBSS	**2.40**	**4.89**	**2.35**	**4.88**	**1.99**	**1.77**
Airport	MBSS	3.61	4.81	3.52	4.76	1.61	1.20
	IP-MBSS	**3.71**	**4.97**	**3.63**	**4.91**	**1.96**	**1.75**
Street	MBSS	4.24	5.00	4.17	4.89	1.74	1.00
	IP-MBSS	**4.42**	**5.56**	**4.39**	**5.38**	**1.99**	**1.87**
Exhibition	MBSS	3.65	4.72	3.60	4.64	2.22	1.29
	IP-MBSS	**3.92**	**4.59**	**3.91**	**4.52**	**2.16**	**2.08**
White	MBSS	5.09	6.87	5.03	6.85	2.69	2.23
	IP-MBSS	**5.25**	**6.86**	**5.23**	**6.86**	**2.01**	**1.83**

Table 4.4.

Objective Evaluation and Comparison of the Proposed Algorithm in Terms of SNR Improvement (dB), and SegSNR Improvement (dB).

Noise Type	Enhancement Algorithms	SNR Improvement (dB)		SegSNR Improvement (dB)	
		0dB	5dB	0dB	5dB
Car	MBSS	4.26	1.01	4.19	0.98
	IP-MBSS	**4.50**	**1.11**	**4.46**	**1.10**
Train	MBSS	3.47	0.82	3.42	0.75
	IP-MBSS	**3.57**	**0.96**	**3.54**	**0.92**
Restaurant	MBSS	2.15	-0.04	2.10	-0.46
	IP-MBSS	**2.27**	**0.04**	**2.24**	**-0.01**
Babble	MBSS	2.27	-0.36	2.21	-0.37
	IP-MBSS	**2.40**	**-0.11**	**2.35**	**-0.12**
Airport	MBSS	3.61	-0.19	3.52	-0.24
	IP-MBSS	**3.71**	**-0.03**	**3.63**	**-0.09**
Street	MBSS	4.24	0	4.17	-0.11
	IP-MBSS	**4.42**	**0.56**	**4.39**	**0.38**
Exhibition	MBSS	3.65	-0.28	3.60	-0.36
	IP-MBSS	**3.92**	**-0.41**	**3.91**	**-0.48**
White	MBSS	5.09	1.87	5.03	1.85
	IP-MBSS	**5.25**	**1.86**	**5.23**	**1.86**

TABLE 4.5.

RESULTS OF NOISE REDUCTION SPEECH QUALITY TEST.

Noise Type	**Enhancement Algorithms**	**PESQ Score**		**MOS Score**	
		0dB	**5dB**	**0dB**	**5dB**
Car	**MBSS**	1.615	1.776	1.8	2.7
	IP-MBSS	**1.693**	**1.915**	**2**	**2.8**
Train	**MBSS**	1.608	1.886	2.6	3.3
	IP-MBSS	**1.693**	**1.893**	**2.3**	**2.9**
Restaurant	**MBSS**	1.697	1.885	1.8	2.7
	IP-MBSS	**1.787**	**1.927**	**1.9**	**2.7**
Babble	**MBSS**	1.665	1.907	1.6	2.7
	IP-MBSS	**1.667**	**2.036**	**1.8**	**2.7**
Airport	**MBSS**	1.774	1.953	1.8	2.8
	IP-MBSS	**1.876**	**2.061**	**1.6**	**2.1**
Street	**MBSS**	1.416	1.866	1.8	2.6
	IP-MBSS	**1.614**	**1.956**	**2**	**2.7**
Exhibition	**MBSS**	1.298	1.633	1.8	2.7
	IP-MBSS	**1.379**	**1.782**	**1.9**	**2.6**
White	**MBSS**	1.433	1.669	2.6	3.5
	IP-MBSS	**1.602**	**1.901**	**2.9**	**3.6**

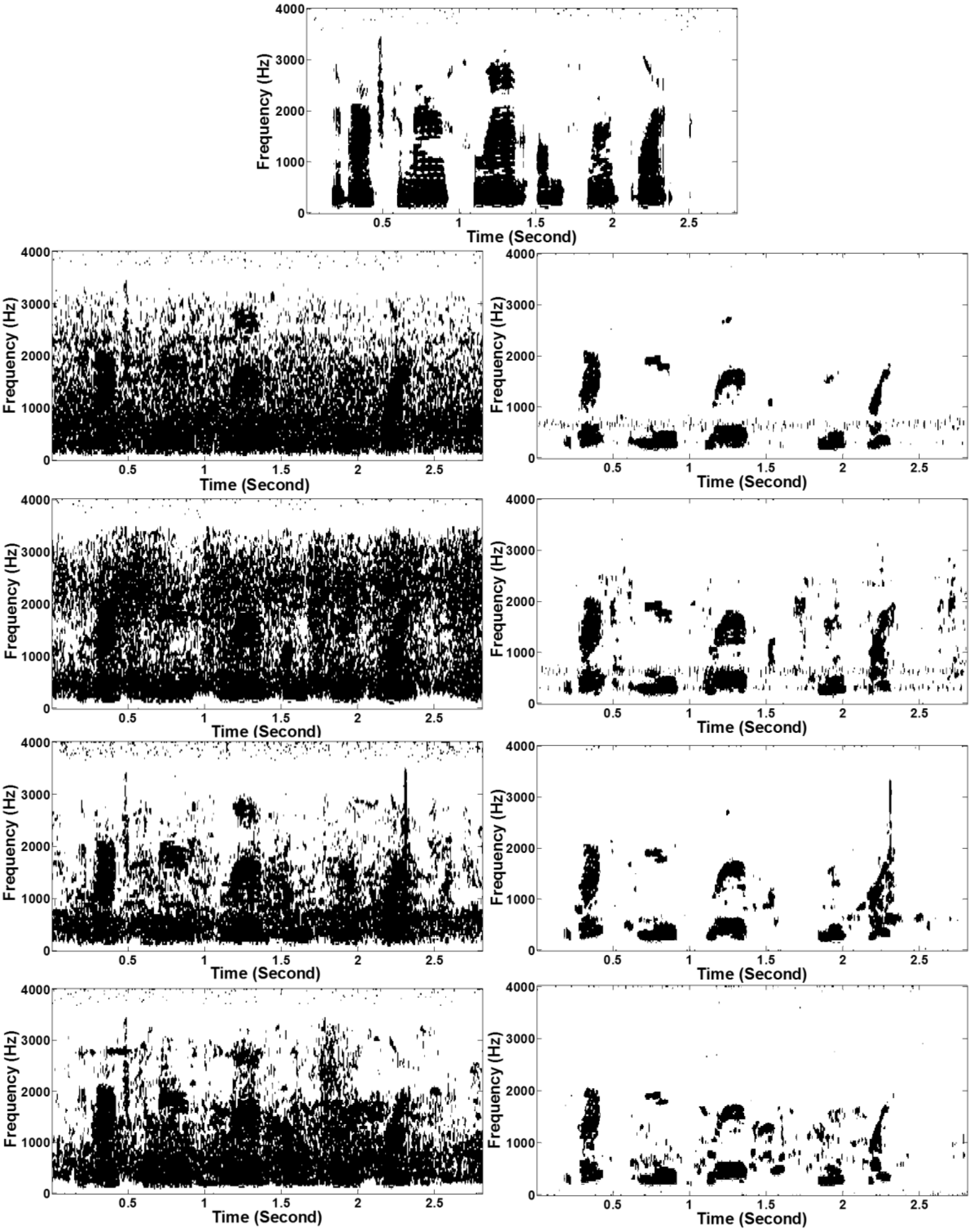


Fig.4.6 (I): Speech spectrogram of *sp1.wav* utterance, "The birch canoe slid on the smooth planks", by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech; (b, d, f, h) speech degraded by car noise, train noise, babble noise, and restaurant noise, respectively (5 dB SNR); (c, e, g, i) corresponding enhanced speech.

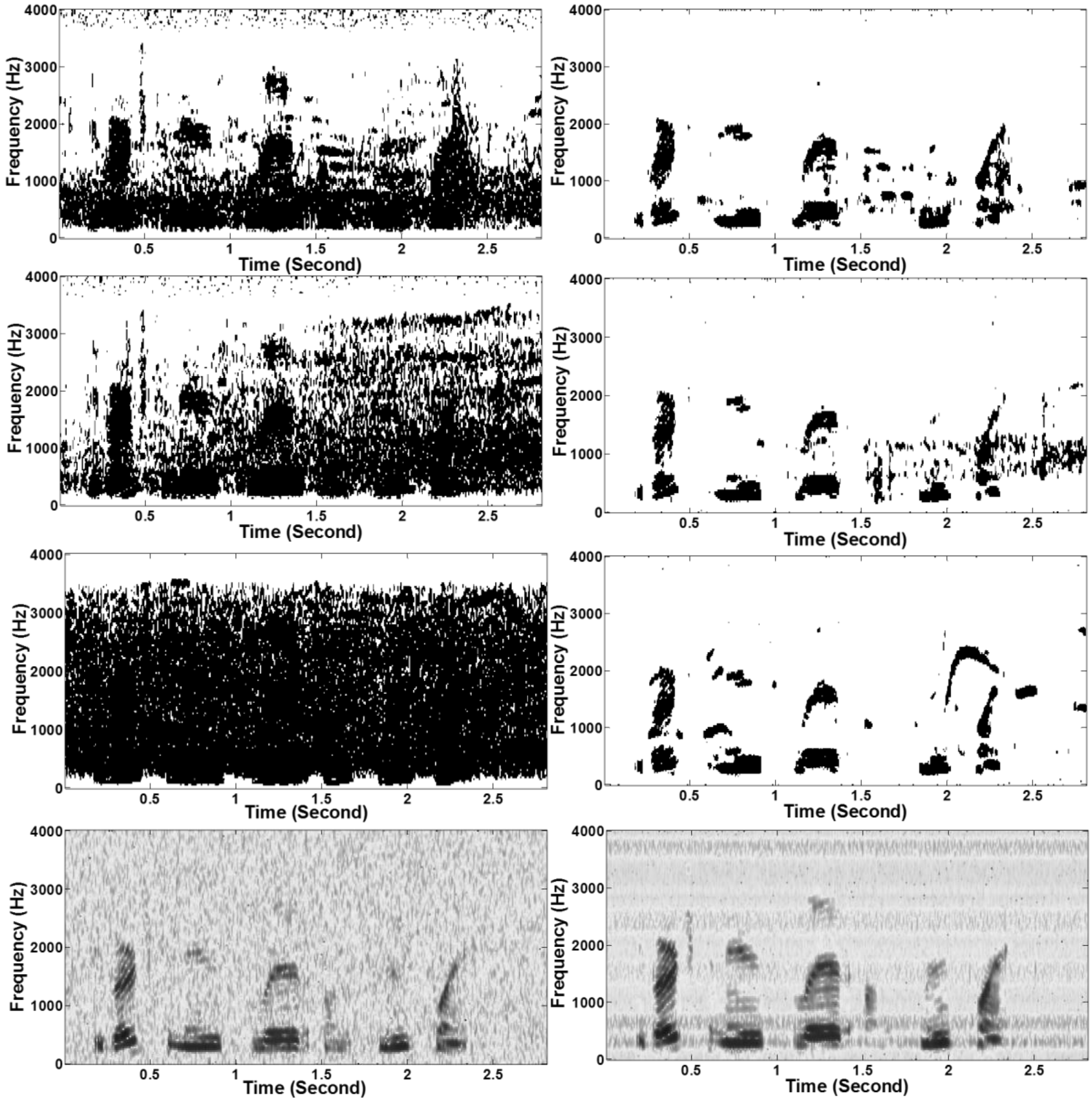


Fig.4.6 (II): Speech spectrogram of *sp1.wav* utterance, "The birch canoe slid on the smooth planks", by a male speaker from the NOIZEUS corpus (From top to bottom): (j, l, n, p) speech degraded by airport, street, exhibition, and white noise, respectively (5 dB SNR); (k, m, o, q) corresponding enhanced speech.

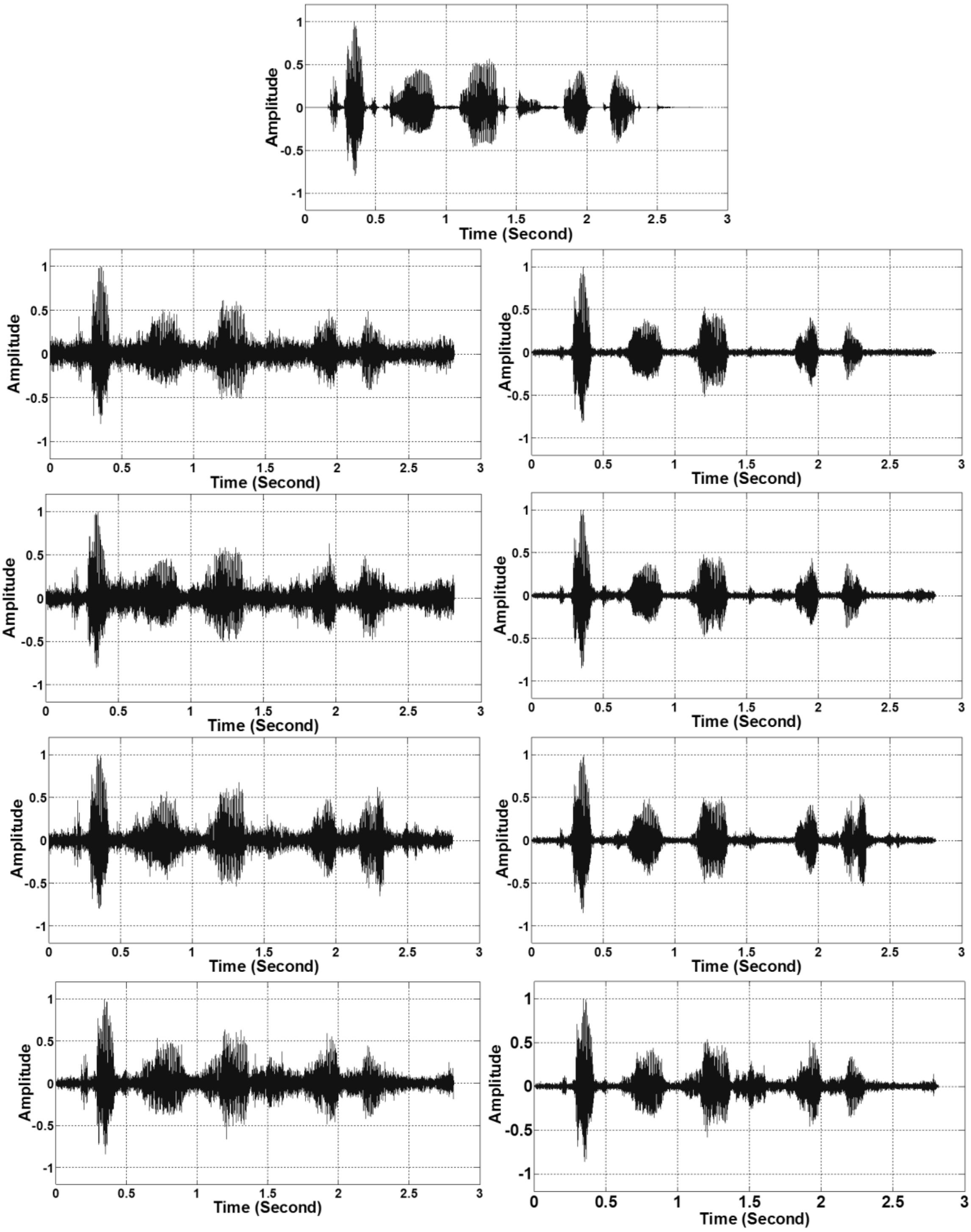


Fig. 4.7(I): Temporal waveforms of *sp1.wav* utterance, "The birch canoe slid on the smooth planks", by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech; (b, d, f, h) speech degraded by car noise, train noise, babble noise, and restaurant noise, respectively (5 dB SNR); (c, e, g, i) corresponding enhanced speech.

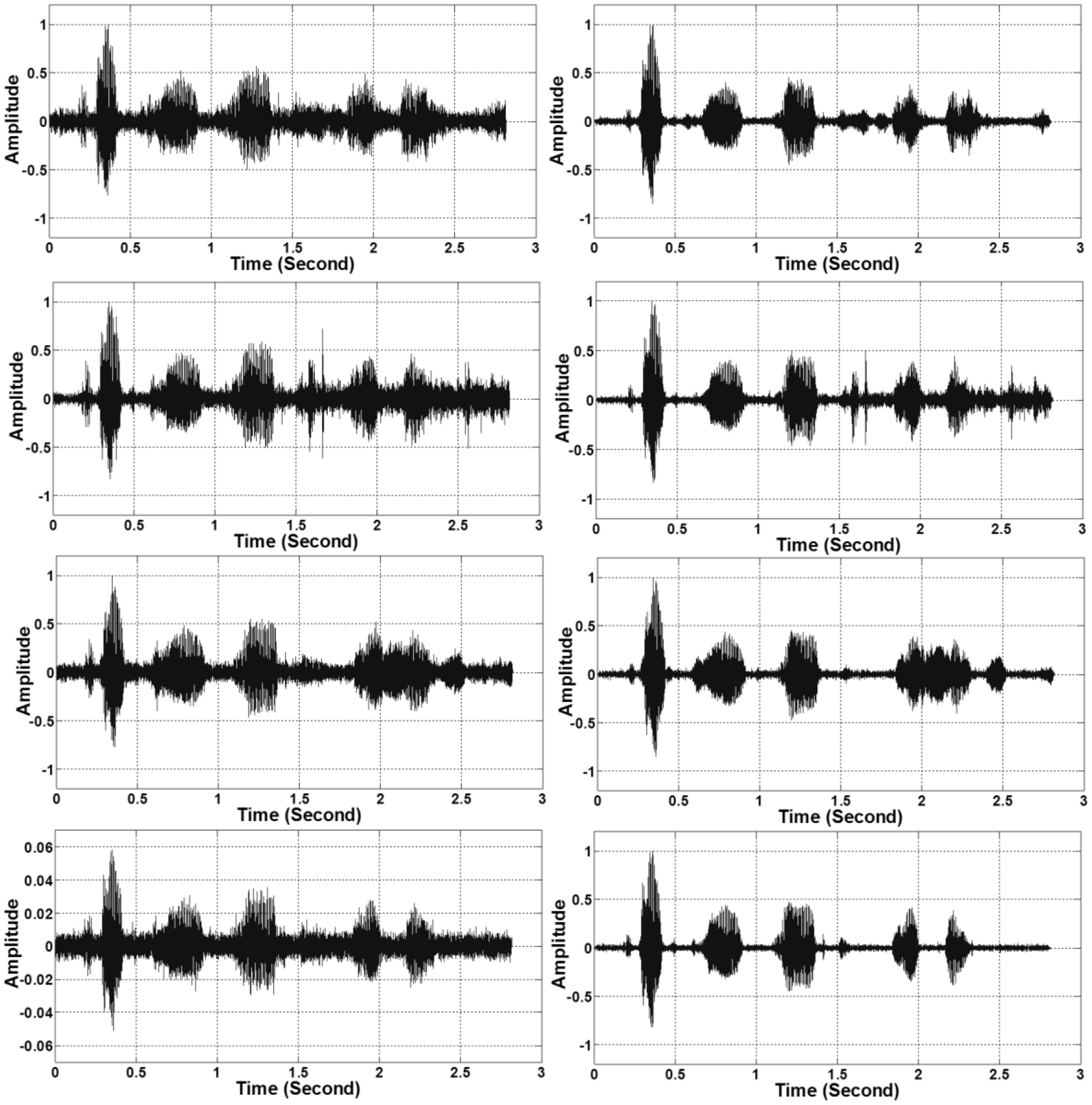


Fig. 4.7(II): Temporal waveforms of *sp1.wav* utterance, "The birch canoe slid on the smooth planks”, by a male speaker from the NOIZEUS corpus (From top to bottom): (j, l, n, p) speech degraded by airport noise, street noise, exhibition noise, and white noise, respectively (5 dB SNR); (k, m, o, q) corresponding enhanced speech.

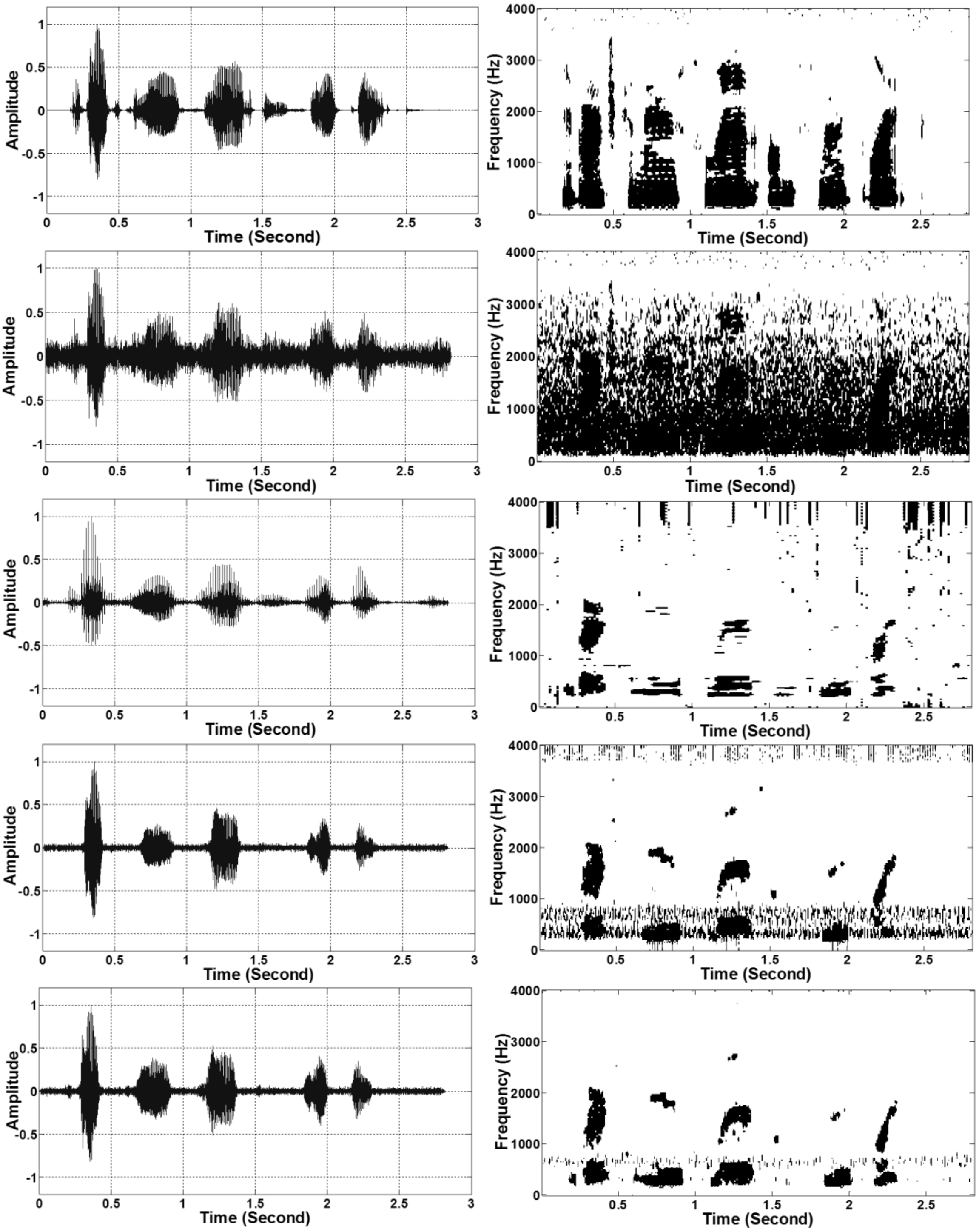


Fig.4.8: Temporal waveforms and speech spectrogram with *sp1.wav* utterance, "The birch canoe slid on the smooth planks", by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech (PESQ = 4.5); (b) noisy speech (degraded by car noise at 5 dB SNR) (PESQ = 1.922); (c) speech enhanced by BSS algorithm (PESQ = 1.9); (c) speech enhanced by MBSS (PESQ = 1.776), and (d) speech enhanced by IP-MBSS (PESQ = 1.915).

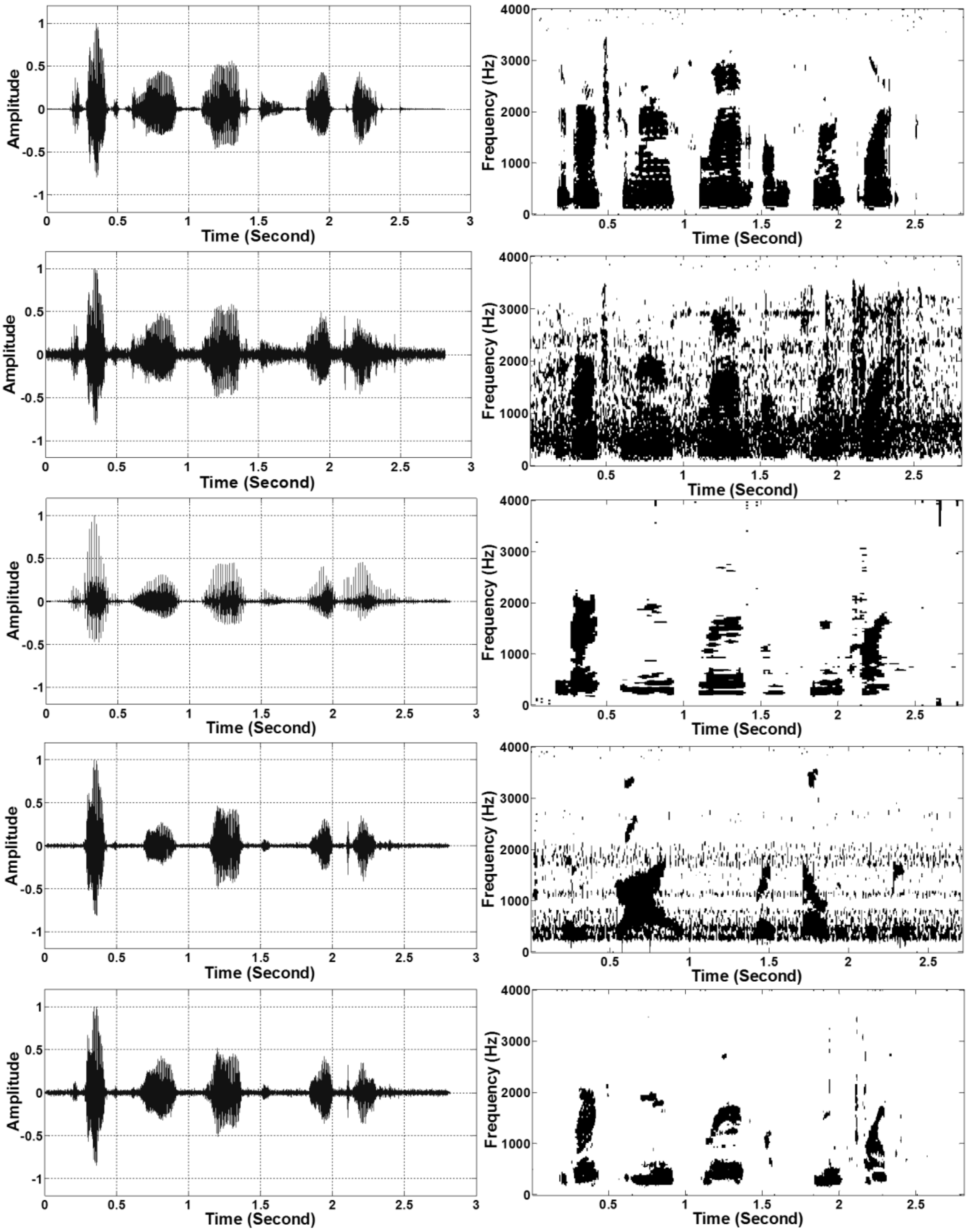


Fig. 4.9: Temporal waveforms and speech spectrogram of *sp1.wav* utterance," The birch canoe slid on the smooth planks", by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech (PESQ = 4.5); (b) noisy speech (degraded by car noise at 10 dB SNR) (PESQ = 2.084); (c) speech enhanced by BSS algorithm (PESQ = 1.898); (d) speech enhanced by MBSS (PESQ = 2.030), and (e) speech enhanced by IP-MBSS (PESQ = 2.147).

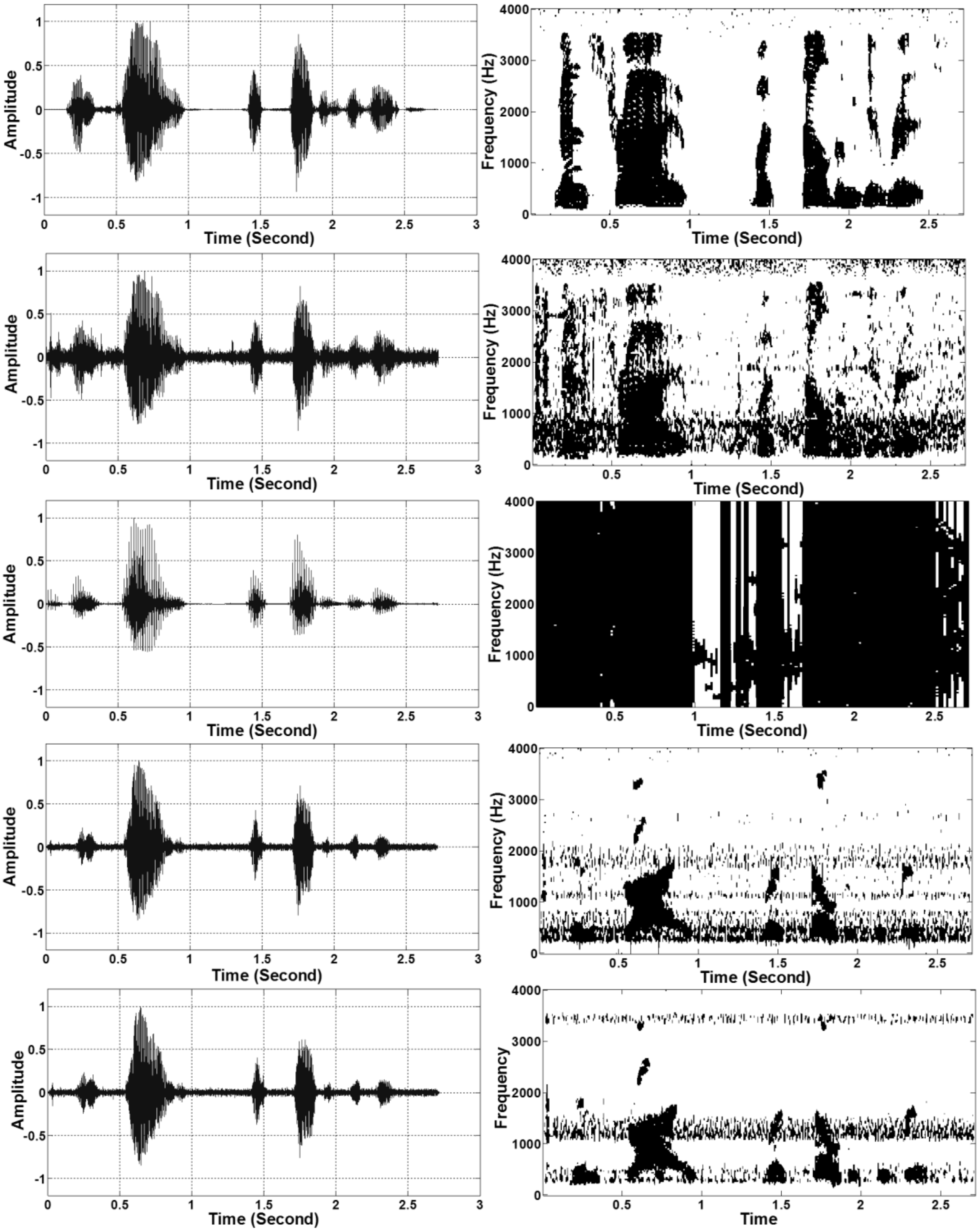


Fig. 4.10: Temporal waveforms and speech spectrogram of *sp 6.wav* utterance, "Men strive but seldom get rich", by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech (PESQ = 4.5); (b) noisy speech (speech degraded by car noise at 10 dB SNR) (PESQ = 2.205); (c) speech enhanced by BSS algorithm (PESQ = 2.231); (d) speech enhanced by MBSS algorithm (PESQ = 2.157); and (e) speech enhanced by IP-MBSS (PESQ = 2.267).

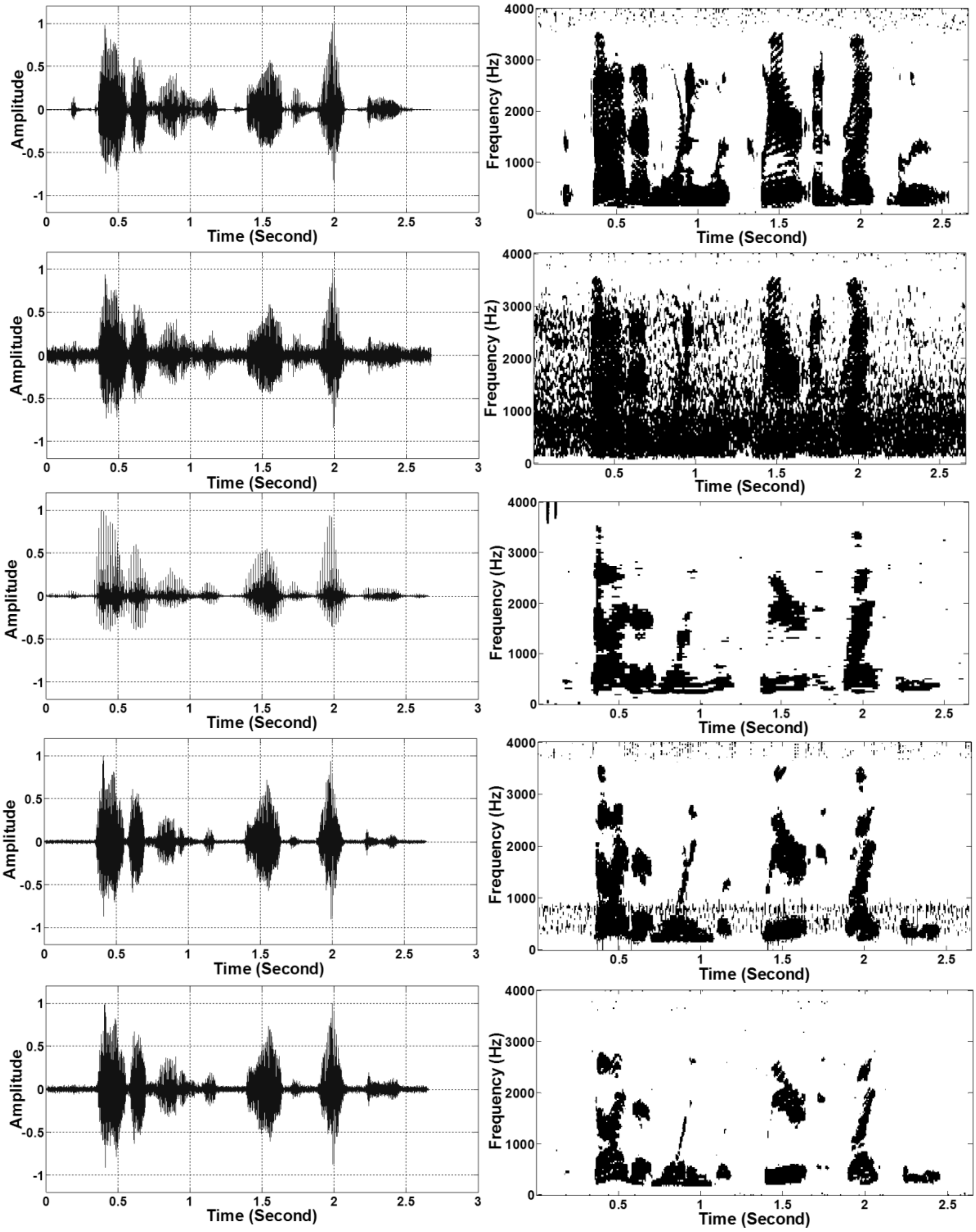


Fig. 4.11: Temporal waveforms and speech spectrogram of *sp10.wav* utterance, "The sky that morning was clear and bright blue", by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech (PESQ = 4.5); (b) noisy speech (speech degraded by car noise at 10 dB SNR) (PESQ = 2.169); (c)speech enhanced by BSS algorithm (PESQ = 2.154); (d) speech enhanced by MBSS (2.259); and (e) speech enhanced by IP-MBSS (2.459).

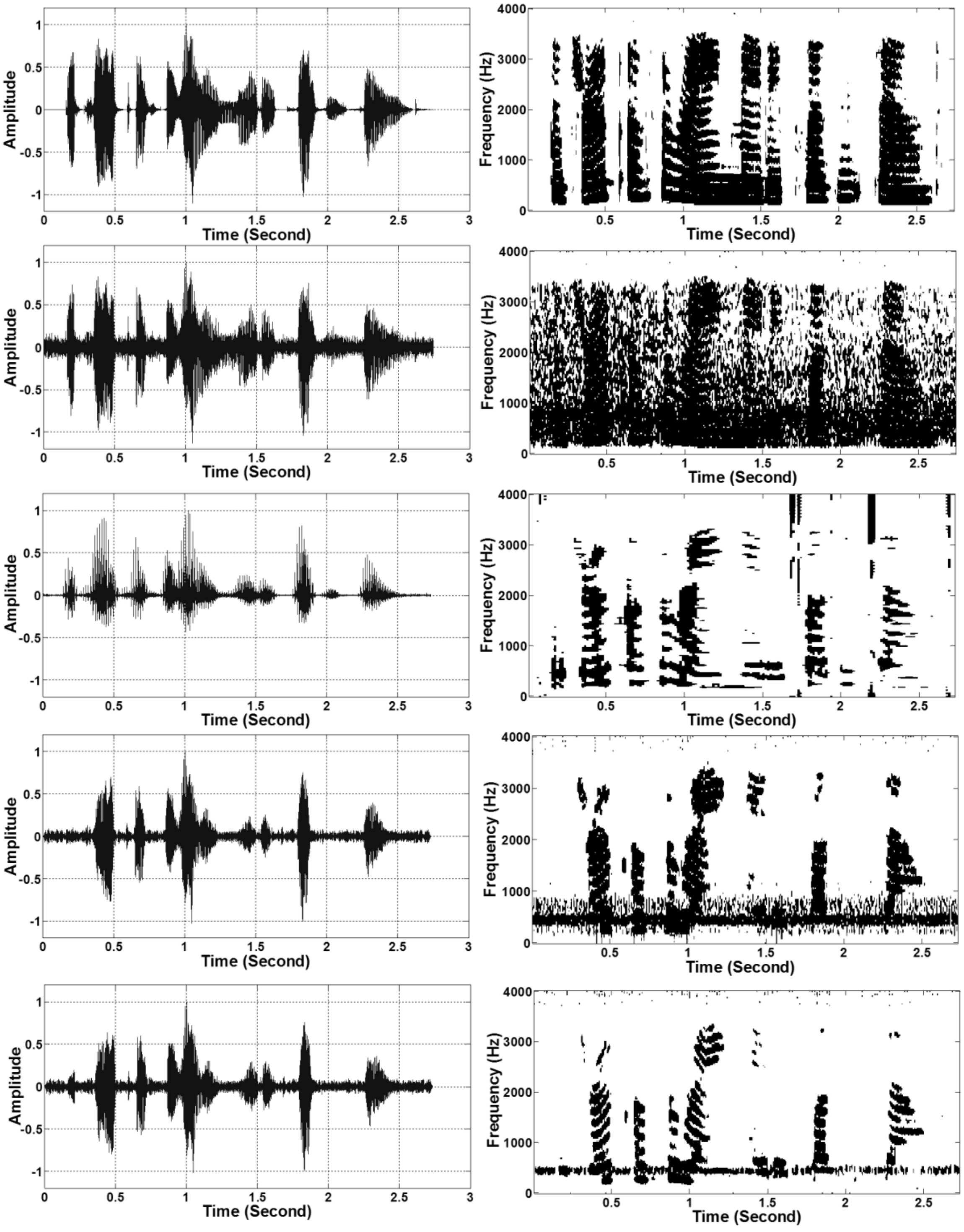


Fig.4.12: Temporal waveforms and speech spectrogram of *sp12.wav* utterance, "The drip of the rain made a pleasant sound", by a female speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech (PESQ = 4.5); (b) noisy speech (degraded by car noise at 10 dB SNR) (PESQ = 2.043); (c) speech enhanced by BSS algorithm (PESQ = 1.782); (d) speech enhanced by MBSS algorithm (PESQ = 2.005); and (e) speech enhanced by IP-MBSS (PESQ = 2.255).

4.6. Summary

In this chapter, an iterative processing based multi-band spectral subtraction (IP-MBSS) algorithm is proposed for the enhancement of speech degraded by non-stationary or colored noises. In the proposed algorithm, IP-MBSS, the output of multi-band spectral subtraction algorithm is used as the input signal again for the next iteration process. The iteration is performed to a limited number of times. After the execution of the reference MBSS algorithm, the additive noise changes to remnant musical noise. The remnant noise is re-estimated at each iteration and the spectral over-subtraction is executed separately, in each band. A comparison with the MBSS algorithm is carried out to evaluate the performance of the proposed algorithm.

Furthermore, the simulation results with different types of noises, have shown that the proposed algorithm, IP-MBSS, with appropriate iteration number reduces the remnant noise tones efficiently that appear in the case of MBSS algorithm and improves the quality and intelligibility of the enhanced speech. The IP-MBSS algorithm is found to perform mostly well for all types of noises at low SNR (≤ 5 dB) except for the case of exhibition noise, in terms of various objective measures including PESQ. It is also evident from the subjective listening tests that the speech enhanced by IP-MBSS algorithm does contains a little amount of remnant noise and speech distortion. Moreover, the remaining remnant noise is of perceptually white quality and the distortions stay within acceptable limit.

In next chapter, an improved multi-band spectral subtraction algorithm is proposed which is based on critical band rate scale of human auditory system and the noise estimation is done in an adaptive manner.

Chapter 5

An Improved Multi-Band Spectral Subtraction based on Critical Band Rate Scale

5.1. Introduction

In the last chapter, we have presented an iterative processing based multi-band spectral subtraction algorithm in which bands are continuous and uniformly frequency spaced. With this algorithm, for various noise types, we have found substantial reduction in remnant musical noise. But, ultimately the humans are the final judge to evaluate the speech quality and intelligibility. Many researchers have applied the perceptual frequency scale of hearing in numerous speech applications [30, 32, 35, 88] for narrowband and wideband speech. Therefore, it is expected that, the processing of bands in accordance to the bands of human hearing in multi-band speech processing will be beneficial in terms of performance. In this chapter we have applied non-uniformly spaced frequency bands that closely match with the perceptual frequency scale of human auditory system for multi-band spectral subtraction algorithm.

The work reported in this chapter has resulted in the following publications

[I]. Navneet Upadhyay, and Abhijit Karmakar, "A perceptually motivated multi-band spectral subtraction algorithm for enhancement of degraded speech," in *Proceedings of 3rd IEEE International Conference on Computer, and Communication Technology*, MNNIT Alllahabad, Nov. 23─25, 2012, pp. 340─345.

[II]. Navneet Upadhyay, and Abhijit Karmakar, "An auditory perception based improved multi-band spectral subtraction algorithm for enhancement of speech degraded by non-stationary noises," in *Proceedings of 4th IEEE International Conference on Intelligent Human Computer Interaction*, IIT Kharagpur, India, Dec. 27─29, 2012, pp. 472─478.

[III]. Navneet Upadhyay, and Abhijit Karmakar, "Single-channel speech enhancement using critical-band rate scale based improved multi-band spectral subtraction," *Journal of Signal and Information Processing*, 2013. [Accepted, Under Print]

This chapter proposes an auditory perception based improved multi-band spectral subtraction (API-MBSS) algorithm for enhancement of speech degraded by non-stationary or colored noise. In the proposed scheme, the whole speech spectrum is divided in different non-uniform frequency spaced bands in accordance to the critical band rate scale and spectral over-subtraction is executed independently, in each band. The proposed algorithm uses an adaptive approach to estimate the noise power from each band without the need of explicit speech pause detection. The noise estimate is updated by adaptively smoothing the noisy signal power in each band. The smoothing parameter is controlled by a linear function of *a-posteriori* signal-to-noise ratio (SNR). This noise estimation approach gives accurate results even at very low SNRs and works continuously, even in the presence of speech. The simulation results as well as evaluations from the objective tests and subjective listening tests show that the proposed algorithm suppresses the noise efficiently and the enhanced speech contains minimal speech distortions with improved SNR.

The rest of this chapter is structured as follows. In Section 5.2, an adaptive noise estimation approach is described which is utilized in the API-MBSS algorithm. In Section 5.3, the proposed algorithm, an improved multi-band spectral subtraction based on critical band rate scale, API-MBSS is presented. Section 5.4, elaborates the experimental results and performance evaluation (subjective and objective measure) and finally Section 5.5 concludes this chapter.

5.2. Noise Estimation

In real-world listening environment, the speech signal is not affected uniformly over the entire frequency spectrum. Some of the frequency components of speech are affected more adversely than others. This kind of noise is referred as non-stationary or colored noise [3].

The noise spectrum estimation is the fundamental component of speech enhancement algorithms. If the noise estimate is too low, annoying remnant noise will be audible, while if the noise estimate is too high, speech will be distorted, resulting possibly in intelligibility loss. There are many approaches to estimate the noise power, especially during speech activity. The non-stationary noise power can be estimated using minimal-tracking algorithms [45, 78, 79], and time-recursive averaging algorithms [46, 79-81].

The minimal-tracking algorithms are based on tracking the minimum of the noisy speech over a finite window. As the minimum is typically smaller than the mean, unbiased estimates of noise spectrum were computed by introducing a bias factor based on the statistics of the minimum estimates. The main drawback of this method is that it takes slightly more than the duration of the minimum-search window to update the noise spectrum when the noise floor increases abruptly. In the recursive averaging type of algorithms [46, 79-81], the noise spectrum is estimated as a weighted average of the past noise estimates and the present noisy speech spectrum. The weights change adaptively depending on the effective SNR of each frequency bin. In this chapter, the non-stationary noise estimate is updated by adaptively smoothing the noisy signal power as a sum of the past noise power and the present noisy signal power without the need of an explicit speech pause detection. Moreover, the smoothing parameter is controlled by a linear function of *a-posteriori* SNR.

In our proposed algorithm, we have estimated and updated the noise spectrum in each frequency band, individually. Reproducing (2.10) for estimation of noise as a first order recursive equation, we obtain

$$|\widehat{D}(\omega,k)|^2 = \lambda(\omega,k)\,|\widehat{D}(\omega,k-1)|^2 + (1-\lambda(\omega,k))|Y(\omega,k)|^2 \tag{5.1}$$

where k is the frame index at frequency ω , $|\widehat{D}(\omega,k)|^2$ is the noise power estimation (i.e. average noise power spectral density, $|\widehat{D}(\omega,k)|^2 = E[|D(\omega,k)|^2]$), in the ω^{th} frequency bin of current frame index k and $|Y(\omega,k)|^2$ is the short-time power spectrum of noisy speech. Further, $\lambda(\omega,k)$ is a time and frequency dependent smoothing parameter whose value depends on the noise changing rate.

The smoothing parameter is the time-varying frequency dependent parameter that is adjusted by the speech presence probability. In [3, 82], the smoothing parameter $\lambda(\omega,k)$ at frame k is selected as a sigmoid function changing with the estimate of the *a-posteriori* signal-to-noise ratio at frame k

$$\lambda(\omega, k) = \frac{1}{1+\exp\left(-a(\mathrm{SNR}\,(\omega,k)-T)\right)} \tag{5.2}$$

where parameter a in sigmoid function (5.2) affects the noise changing rate and is a constant with a value between 1 to 6. The parameter T in (5.2) is the center offset of the transition curve in sigmoid function and the value of T is around 3 to 5. A plot of smoothing parameter against the *a-posteriori* SNR at different values of a and different values of T is shown in Fig. 5.1 (i) and (ii), respectively. These curves are obtained with experimental conditions given in Table 5.3 and explained in Section 5.4.

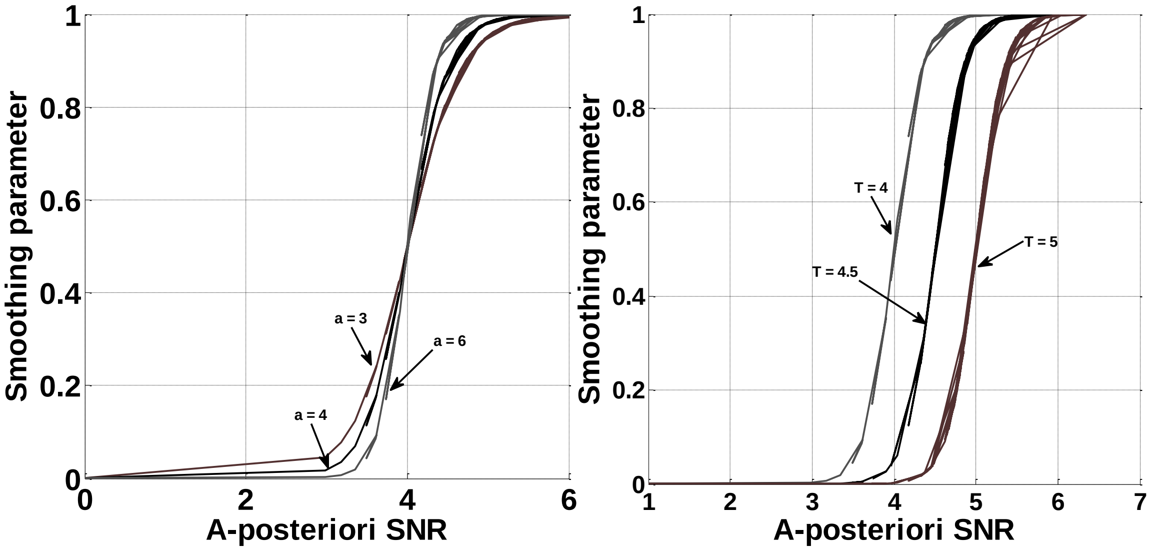


Fig. 5.1: Plot of smoothing parameter against the *a-posteriori* SNR: i) for different values of a, and ii) for different values of T.

It is also to be noted that, the smoothing function has also been obtained differently. In [3, 82], a different function was proposed for computing $\lambda(\omega, k)$ as

$$\lambda(\omega, k) = 1 - \min\left\{1, \frac{1}{\left(\mathrm{SNR}(\omega,k)\right)^p}\right\} \tag{5.3}$$

where p is an integer, and $\mathrm{SNR}(\omega, k)$ is given by (5.4).

The updation of noise estimate is performed on a continuous manner. This is accomplished by controlling the smoothing factor $\lambda(\omega,k)$ depending on the *a-posteriori* signal-to-noise ratio at frame k , defined as

$$\mathrm{SNR}(\omega,k) = 10\log_{10}\left(\frac{|Y(\omega,k)|^2}{\frac{1}{m}\sum_{p=1}^{m}|\hat{D}(\omega,k-p)|^2}\right) \tag{5.4}$$

where, the denominator part is the average of the noise estimate of the previous m frames (value of m usually varying from 5 to 10).

The slope parameter a in (5.2) controls the way in which smoothing parameter $\lambda(\omega,k)$ changes with *a-posteriori* SNR. Generally, larger values of a in (5.2) lead to larger values of $\lambda(\omega,k)$ and slower noise updates, whereas smaller values of a in (5.2) give faster noise updates, at the risk of possible over-estimation during long voiced intervals. It results in smoothing parameter being close to 0 when the speech is absent in frame k, that is, the estimate of noise power in frame k follows rapidly the power of the noisy signal in the absence of speech. On the other hand, if speech signal is present, the new noisy signal power is much larger than the previous noise estimate. Therefore, the value of smoothing parameter increases rapidly with increasing the value of SNR. Hence, the noise update is slower or eventually stops because of the larger value of the smoothing parameter, as shown in Fig. 5.4. Theoretically, the *a-posteriori* SNR should always be 1 when noise alone is present and greater than 1 when both speech and noise are present.

The main advantage of using the time-varying smoothing factor $\lambda(\omega,k)$, is that the noise power can be adapted differently, at different rates in the various frequency bins, and depends on the estimate of the *a-posteriori* signal-to-noise ratio at frame k in that bin. In the next section, the auditory perception based improved multi-band spectral subtraction (API-MBSS) algorithm is elaborated which utilizes this noise estimation approach.

5.3. An Auditory Perception based Improved Multi-Band Spectral Subtraction Algorithm

It is well-known that the sensitivity of human ear varies non-linearly in the frequency spectrum [84]. Therefore, the notion of critical band (CB) is important for describing hearing sensations such as perception of loudness, pitch, and timbre. A commonly used scale for this purpose is the Bark scale or critical band rate scale. Theoretically, the range of human auditory frequency spreads from 20 Hz to 20 kHz and covers approximately 24 CBs. However, the frequency range of the narrowband human voice is typically only from about 300 Hz to 3.4 kHz. The bands in the proposed algorithm are derived in a manner that it closely matches the psychoacoustic frequency scale of human ear , i.e., the critical band rate scale.

Based on the measurements by Zwicker *et* al. [84, 85], the critical band rate scale can approximately be expressed in terms of the linear frequency as

$$z(f) = 13\tan^{-1}(7.6\times10^{-4}\, f) + 3.5\tan^{-1}(1.33\times10^{-4}\, f)^2 \quad \text{[Bark]} \tag{5.5}$$

Here, $z(f)$ is the CB rate scale in Bark, and f is the physical frequency in Hz. The underlying sampling rate was chosen to be 8 kHz for implementing the proposed algorithm. Fig. 5.2 shows a mapping between the physical (linear) frequency scale and the CB rate scale [85]. The corresponding critical bandwidth (CBW) of the center frequencies can be expressed by

$$\mathrm{CBW}(f_c) = \; 25 + 75\,(1 + 1.4\times10^{-6} f_c^2)^{0.69} \tag{5.6}$$

where f_c is the center frequency (Hz). Within this bandwidth, there are approximately 18 CBs as listed in Table 5.1 [85]. According to the specifications of center frequencies, lower and upper edge frequencies are given in Table 5.1.

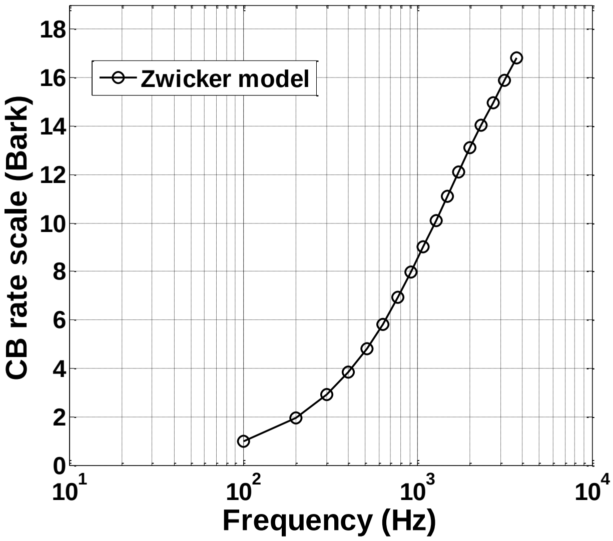


Fig. 5.2: Mapping between the physical frequency scale and critical band rate scale.

Table 5.1.

Critical Bands of the Human Auditory System for Frequency Bandwidth of 4 kHz.

CB rate (Bark)	Lower edge Freq. (Hz)	Upper edge Freq. (Hz)	Center Freq. (Hz)	CBW (Hz)	
1	20	100	50	100	Band 1
2	100	200	150	100	
3	200	300	250	100	
4	300	400	350	100	Band 2
5	400	510	450	110	
6	510	630	570	120	
7	630	770	700	140	Band 3
8	770	920	840	150	
9	920	1080	1000	160	
10	1080	1270	1170	190	Band 4
11	1270	1480	1370	210	
12	1480	1720	1600	240	
13	1720	2000	1850	280	Band 5
14	2000	2320	2150	320	
15	2320	2700	2500	380	
16	2700	3150	2900	450	Band 6
17	3150	3700	3400	550	
18	3700	4000	3850	-	

It is inefficient to separate the whole speech spectrum into such a large number of non-uniformly frequency spaced intervals, as given in Table 5.1, for our proposed algorithm. This is because it is very difficult to set the values of additional band over-subtraction or scale factor empirically, for each band separately. Therefore, after several implementations with various numbers of bands, it has been found that the performance of the algorithm does not improve, for bands numbering more than six. Thus, the critical bands, as in Table 5.1 are grouped together into six non-uniform bands each containing three consecutive critical bands. Therefore, the spectrum analysis is performed in a total number of six non-uniformly spaced frequency bands, matching closely with the the human auditory system. The six frequency bands are with ranges of {20 Hz ～ 300 Hz (Band 1), 300 Hz ～ 630 Hz (Band 2), 630 Hz ～ 1080 Hz (Band 3), 1080 Hz ～ 1720 Hz (Band 4), 1720 Hz ～ 2700 Hz (Band 5), 2700 Hz ～ 4 kHz (Band 6)} and are tabulated in Table 5.2.

TABLE 5.2.
THE CRITICAL BAND RATE SCALE BASED
NON-UNIFORM BANDS FOR API-MBSS.

Bands	**Frequency range**
Band 1	20 Hz - 300 Hz
Band 2	300 Hz - 630 Hz
Band 3	630 Hz - 1080 Hz
Band 4	1080 Hz - 1720 Hz
Band 5	1720 Hz - 2700 Hz
Band 6	2700 Hz - 4000 Hz

Therefore, as in (2.18), the estimate of the clean speech spectrum in the i^{th} Band is obtained by

$$|\hat{S}_i(\omega)|^2 = \begin{cases} |Y_i(\omega)|^2 - \alpha_i.\delta_i.|\widehat{D}_i(\omega)|^2, & \text{if } |\hat{S}_i(\omega)|^2 > \beta.|Y_i(\omega)|^2 \\ \beta.|Y_i(\omega)|^2 & \text{else} \end{cases} \tag{5.7}$$

where $\omega_i < \omega < \omega_{i+1}$,

Here, ω_i and ω_{i+1} are the start and end frequency bins of the i^{th} Band, α_i is the band specific over-subtraction factor. Note that, α_i is a function of the segmental SNR.

In our algorithm, the over-subtraction factor has been modified for the perceptual band specific analysis to obtain the spectral over-subtraction. The segmental SNR_i is computed using spectral components of noisy speech and noise estimate $\sum_{\omega=\omega_i}^{\omega_{i+1}} |Y_i(\omega)|^2$, and $\sum_{\omega=\omega_i}^{\omega_{i+1}} |\widehat{D}_i(\omega)|^2$ for each band $i = 1, 2, \dots, K$. Here, i is the non-uniformly spaced frequency band number, $K = 6$ is the total number of perceptual bands. The segmental SNR of the i^{th} Band can be computed as

$$\mathrm{SNR}_i\ (\mathrm{dB}) = 10 \log_{10}\left(\frac{\sum_{\omega=\omega_i}^{\omega_{i+1}} |Y_i(\omega)|^2}{\sum_{\omega=\omega_i}^{\omega_{i+1}} |\widehat{D}_i(\omega)|^2}\right) \tag{5.8}$$

where $|\widehat{D}_i(\omega)|^2$ is estimated using (5.1). For continuity of description, the expression of the band specific over-subtraction factor is reproduced from Chapter 2 (2.20) (see Fig. 2.3). The band specific over-subtraction can be calculated as

$$\alpha_i = \begin{cases} \alpha_{\max} & \text{if } \mathrm{SNR}_i \le \mathrm{SNR}_{\min} \\ \alpha_{\max} + (\mathrm{SNR}_i - \mathrm{SNR}_{\min})\left(\frac{\alpha_{\min} - \alpha_{\max}}{\mathrm{SNR}_{\max} - \mathrm{SNR}_{\min}}\right), & \text{if } \mathrm{SNR}_{\min} \le \mathrm{SNR}_i \le \mathrm{SNR}_{\max} \\ \alpha_{\min} & \text{if } \mathrm{SNR}_i \ge \mathrm{SNR}_{\max} \end{cases} \tag{5.9}$$

The scale factor δ_i , in (5.7), is used to provide an additional degree of control over the noise subtraction level in each band. The values of δ_i is empirically determined and set to

$$\delta_i = \begin{cases} 0.8 & f_i \le 1\ \text{kHz} \\ 1.3 & 1\ \text{kHz} < f_i \le \frac{f_s}{2} - 2\ \text{kHz} \\ 1 & f_i > \frac{f_s}{2} - 2\ \text{kHz} \end{cases} \tag{5.10}$$

where f_i is the upper-end frequency of i^{th} Band and f_s is the sampling frequency. Since, most of the speech energy is present in the lower frequencies, smaller values of δ_i are used for the low-frequency bands in order to minimize speech distortion.

The factors δ_i and α_i can be adjusted for each critical band for different speech conditions to get better speech quality. Hence the estimate of the clean speech spectrum in the i^{th} Band can be obtained by (5.7). Negative values resulting from the subtraction in (5.7) are floored to the noisy spectrum by setting the maximum attenuation threshold β to 0.03. In Fig. 5.3 the block diagram of complete auditory perception based improved multi-band spectral subtraction algorithm is shown.

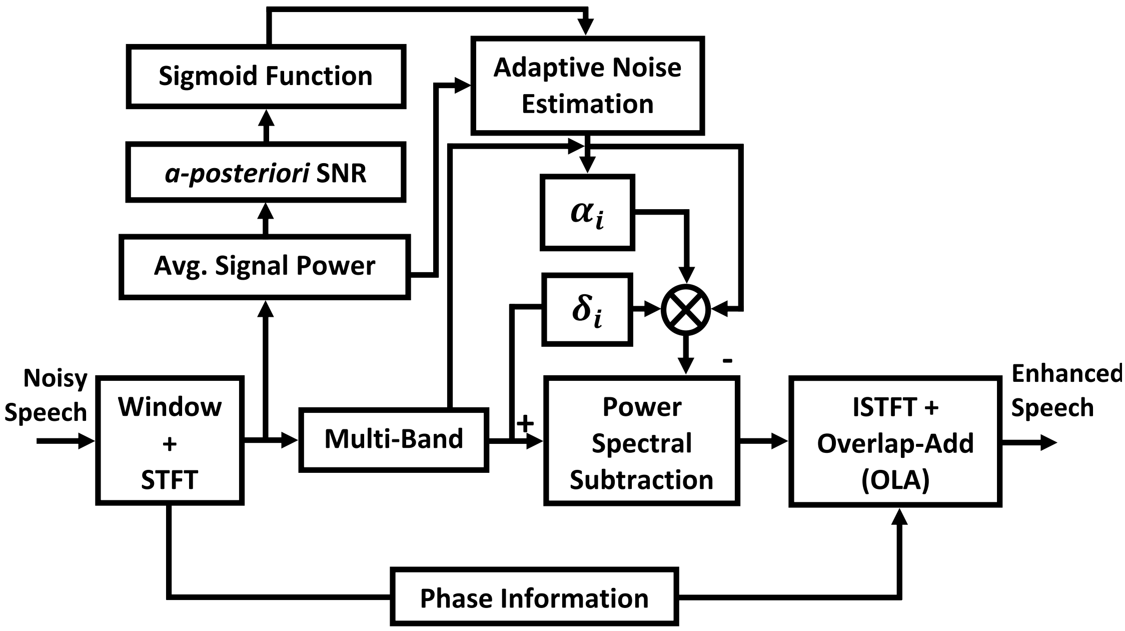


Fig. 5.3: Block diagram of critical band rate scale based improved multi-band spectral subtraction algorithm.

5.4. Experimental Results and Performance Evaluation

This section presents the performance evaluation of the proposed speech enhancement algorithm, API-MBSS, described in this chapter, and its comparison with other subtractive-type algorithms. The noisy speech samples have been taken from NOIZEUS corpus speech database [76]. The NOIZEUS database is composed of 30 phonetically balanced sentences belonging to six speakers, three male and three female, degraded by seven different real-world noises at different SNRs. The corpus is sampled at 8 kHz, quantized linearly using 16 bits and filtered to simulate receiving frequency characteristics of telephone handsets. A total of four different utterances, from NOIZEUS corpus, are used in our evaluation. Every sentence has a silence segment in the beginning that lasts more than 0.25 ms (approx). Noise signals have different time-frequency distributions, and therefore a different impact on speech. Hence, eight types of noises, seven real-world noises, as in NOIZEUS, and a computer generated white Gaussian noise, have been used for the evaluation. For our purpose, the sentences are degraded at varying SNR levels, i.e., at global SNR levels of 0 dB to 15 dB in steps of 5 dB. The real-world noises are car, train, restaurant, babble, airport, street, and exhibition noise. The performance of the proposed speech enhancement algorithm is tested on such noisy speech samples. We have used MATLAB software as the simulation environment. The performance of the API-MBSS has been compared with basic spectral subtraction (BSS), spectral over-subtraction (SOS) and MBSS (with averaging-based and adaptive-based noise estimation) speech enhancement algorithms.

For our enhancement experiments, the 8 kHz sampled noisy speech signals are quantized into digital signals with 16 bit resolution. The frame size is chosen to be 256 samples, i.e., a time frame of 32 ms, with 50% overlapping. The sinusoidal Hamming window with size 256 samples is applied to the noisy signal. The noise estimate is updated adaptively and continuously using the smoothing parameter (5.1). For calculation of smoothing parameter, the value of a and T is chosen to be 4 and 5, respectively in the sigmoid function (5.2). Fig. 5.4 shows the frame-by-frame update of smoothing parameter in different frames and Table 5.3, depicts the experimental conditions used in our implementation.

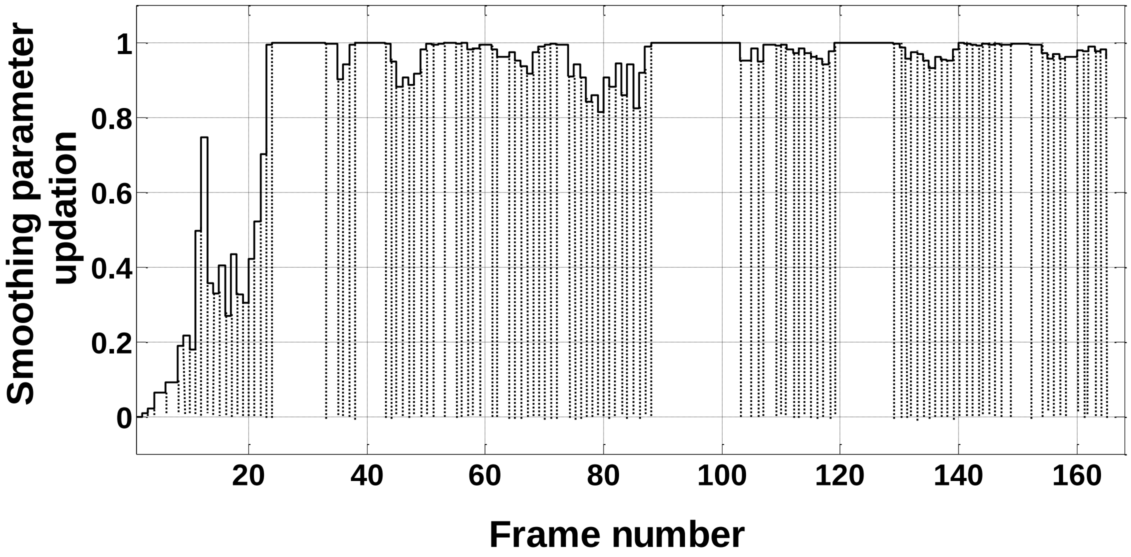


Fig. 5.4: Smoothing parameter updation.

TABLE 5.3.

EXPERIMENTAL CONDITIONS.

Sampling rate	8 kHz or 8000 samples/sec.
Quantization bit rate	16 bit
Frame length	32 ms (256 sample points)
Overlap	16 ms (125 sample points, 50% overlapping frames)
FFT length	256 points FFT
Window function	Hamming window (256 sampless)
Speech Corpus	NOIZEUS
Noises (7 Real-world noise and a computer generated noise)	(Car, Train, Restaurant, Babble, Airport, Street, Exhibition), and (White Gaussian)
Objective evaluation	SNR, SegSNR, ISD, and PESQ
Subjective evaluation	Spectrogram and listening tests (MOS)

For the uniformly frequency spaced multi-band spectral subtraction (MBSS) algorithm, the over-subtraction factor α_i is computed for each frequency band [21]. In this algorithm, four uniformly spaced frequency bands {60 Hz ～ 1 kHz (Band 1), 1 kHz ～ 2 kHz (Band 2), 2 kHz ～ 3 kHz (Band 3), 3 kHz ～ 4 kHz (Band 4)} have been taken. The value of over-subtraction factor α_i is determined using Fig. 2.4 and (2.20), and the value of additional over-subtraction factor δ_i for each band is set as per (2.21). The value of spectral flooring parameter β is taken as 0.03 and noise estimate is updated during the silence frames by using averaging [20].

For the implementation of the proposed auditory perception based improved multi-band spectral subtraction (API-MBSS) algorithm, we compute α_i in accordance to (5.9) and then apply spectral over-subtraction in each band, independently. The number of bands that gives an optimal speech quality is found to be six, and these six non-uniformly frequency spaced bands have been taken as per the critical band rate scale, given in Table 5.2. The noise is estimated by using the adaptive noise estimation approach in each band as given in Section 5.2. The value of δ_i is fixed as per (5.10) and the value of other parameters have been taken to be same as the reference MBSS algorithm [21]. The performance of API-MBSS is evaluated using both objective measures and subjective listening tests.

The input SNR vs. output SNR of API-MBSS in comparison to MBSS algorithm, spectral over-subtraction (SOS) algorithm and basic spectral subtraction (BSS) algorithm for real-world noises and white Gaussian noises has been shown in Table 5.4. The amount of noise reduction is usually measured with the SNR improvement which is given by the difference between input SNR and output SNR. The SNR improvements for BSS, SOS, MBSS, and API-MBSS algorithms are presented in Table 5.5. The SNR measure does not demonstrate much consistency in its performance, as suggested by informal listening tests. Therefore, apart from SNR, three other objective measures, namely the segmental SNR (SegSNR), Itakura-Saito distortion (ISD), and perceptual evaluation of speech quality (PESQ) are used under various noisy environments to evaluate and compare the performance of the proposed speech enhancement algorithm. SegSNR is defined as the average ratio of signal energy to noise energy per frame, and is regarded to be better correlated with perceptual quality than the SNR. In other words, it is well-known that Seg SNR is more accurate in indicating the speech distortion than the overall SNR. The higher value of the SegSNR indicates the weaker speech distortions. The input SegSNR vs. output SegSNR and ISD of API-MBSS over MBSS, and SOS algorithm for real-world noises and computer generated white Gaussian noises has been shown in Table 5.4. The SegSNR improvements for SOS, MBSS, and API-MBSS algorithms are presented in Table 5.5. Fig. 5.5 shows the performance of API-MBSS algorithm at various real-world noises and white Gaussian noise at different SNR levels and Fig. 5.6 shows the performance improvement of API-MBSS at various real-world noises and white Gaussian noise at different SNR levels over SOS and MBSS.

From the results given in Table 5.4, Table 5.5, and Fig. 5.5 - Fig. 5.6, we can conclude that the API-MBSS shows superior results for all types noises (non-stationary and stationary) at low SNR (≤ 10 dB) except for the case of restaurant noise, while for SNR (≈ 15 dB) it shows better results for car, exhibition, white (stationary) noises and gives poorer results for street noise.

Also, this proposed algorithms shows better performance than multi-band spectral subtraction and basic spectral subtraction algorithm at all levels of SNRs. Further, API-MBSS shows comparable performance with spectral over-subtraction algorithm except for street noise case. For the stationary noise (AWGN) case, our results lie in the middle.

Table 5.6 presents the PESQ score results and MOS score results of enhanced speech by API-MBSS in comparison with BSS, SOS and MBSS algorithm. The PESQ score results of enhanced speech by API-MBSS is more for SNR > 5 dB in all noisy conditions except for white noise while MOS score is also more except for car noise, babble noise and white noise cases.

Objective measures do not give indications about the structure of the remnant noise. Speech spectrogram constitutes a well-suited tool for observing this structure. Fig. 5.8 shows the spectrogram obtained with the proposed algorithm for the speech sentence (sp10) degraded by car noise, train noise, babble noise, restaurant noise, airport, street, exhibition, and white noise, respectively at 10 dB SNR.

In Fig. 5.9, the temporal waveforms are shown which are obtained with API-MBSS for the speech sentence (sp10) degraded by car noise, train noise, babble noise, restaurant noise, airport, street, exhibition, and white noise, respectively at 10 dB SNR. Fig. 5.10 - Fig. 5.13 shows the temporal waveforms and spectrogram of the enhanced speech obtained with the proposed algorithm for the speech sentence pronounced by male speaker (sp10, sp 6, sp1), female speaker (sp12) and degraded by car noise at 10 dB SNR in comparison with MBSS, and API-MBSS algorithm where the noise is estimated with averaging-based and adaptive-based noise estimation.

It can be seen from Figs. 5.8 - Fig. 5.13 that the reduction of the musical structure of the remnant noise by the proposed algorithm is more pronounced compared to MBSS algorithm. Therefore, speech enhanced with the proposed algorithm is more pleasant and the remnant noise has a "*perceptually white quality*" while distortion remains acceptable. Also, the informal listening tests indicate that the API-MBSS not only reduces the low-frequency noise, but also eliminates the high-frequency noise substantially. This confirms the values of the objective measures given in Table 5.4, Table 5.5 and it is validated by listening tests, given in Table 5.6.

Subjective tests are also performed for comparison with different subtractive-type algorithms. These tests confirm that the proposed enhancement algorithm leads to the better result for a human listener compared to other subtractive-type algorithms. Fig. 5.7 shows a scatter plot of MOS vs. PESQ scores of API-MBSS for various types of real-world noises and a computer generated white Gaussian noise. In Fig 5.7, a straight line with a constant slope is provided as a reference. It is observed that the PESQ scores are better correlated with MOS scores for airport, street, and car noise in comparison to train and babble noises for most of the conditions.

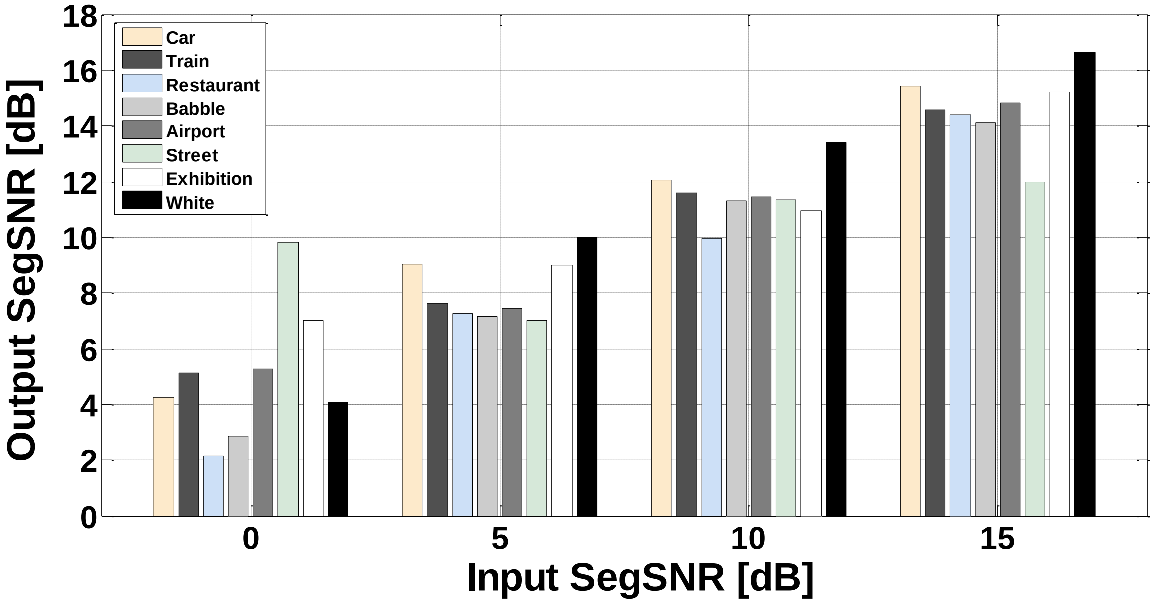


Fig. 5.5: Output SegSNR of API-MBSS for car, train, restaurant, babble, airport, street, exhibition, and white Gaussian noises.

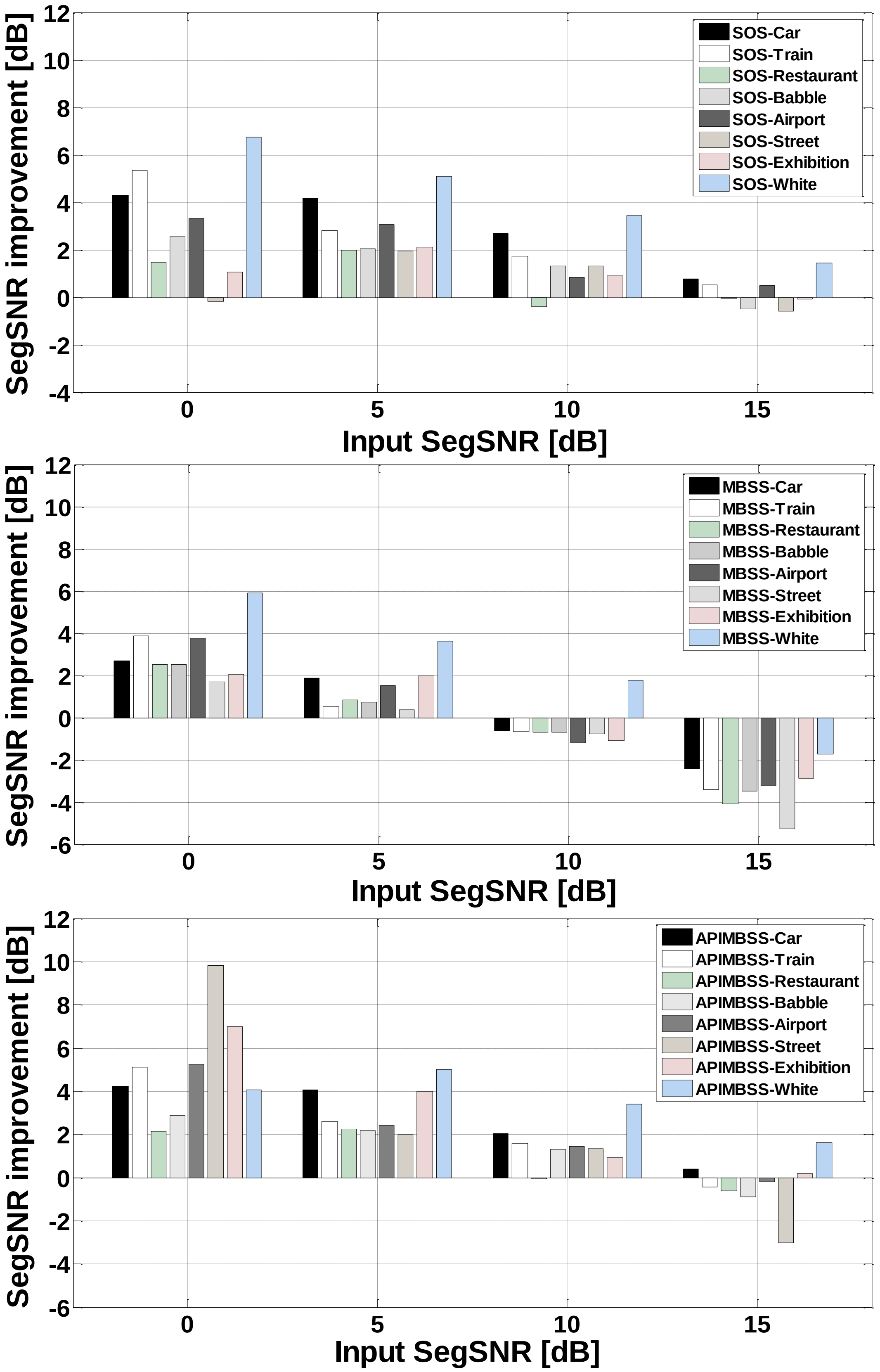


Fig. 5.6: SegSNR improvement of API-MBSS over SOS and MBSS for car, train, restaurant, babble, airport, street, exhibition, and white Gaussian noises.

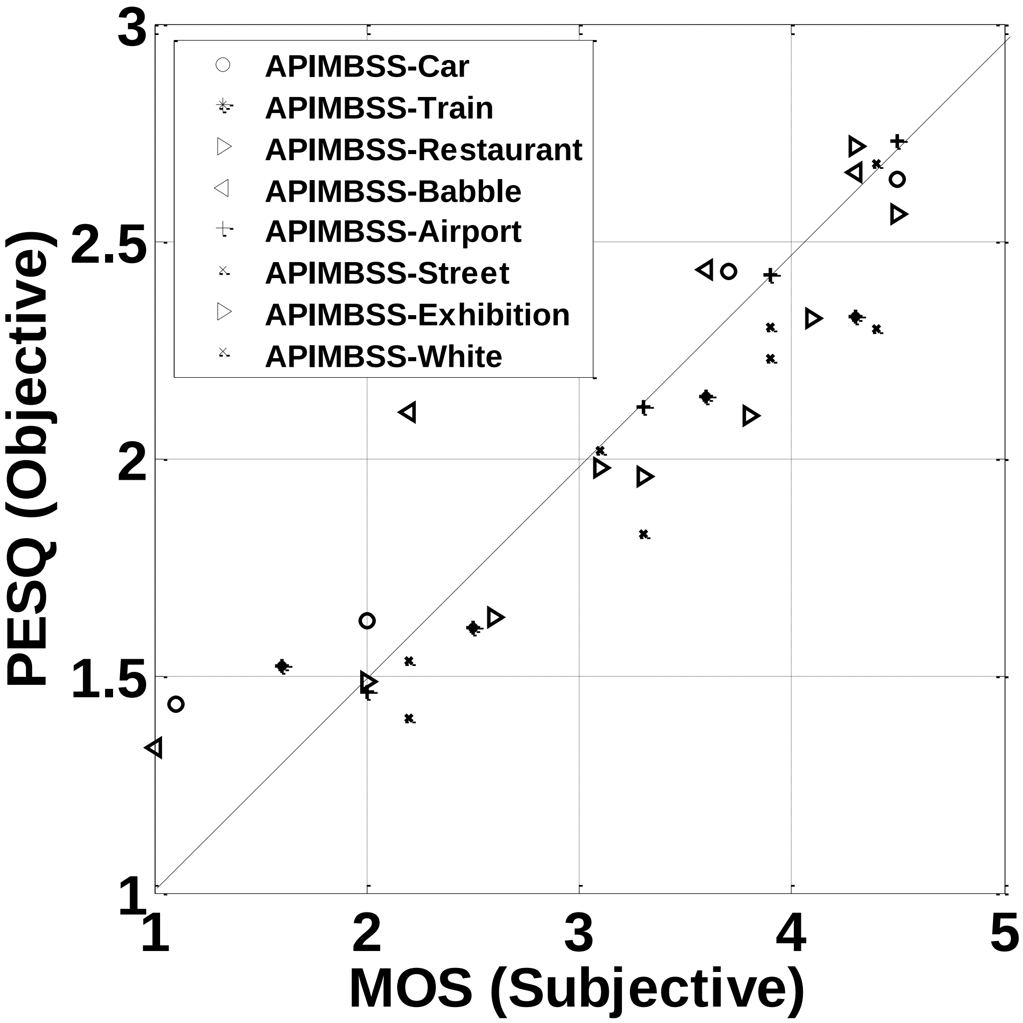


Fig. 5.7: Scattered plot of PESQ score vs. mean MOS score of API-MBSS for car, babble, restaurant, train, airport, street, exhibition, and white noises.

TABLE 5.4.

OUTPUT SNR (GLOBAL), OUTPUT SEGSNR AND ITAKURA–SAITO DISTANCE (ISD) MEASURE RESULTS OF ENHANCED SPEECH SIGNALS AT (0, 5, 10, 15) DB SNRS. ENGLISH SENTENCE *sp10.wav* PRODUCED BY A MALE SPEAKER IS USED AS ORIGINAL SPEECH SIGNAL.

Noise Type	Enhancement Algorithms	SNR (dB)				SegSNR (dB)				ISD			
		0 dB	**5 dB**	**10 dB**	**15 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**
Car	**BSS**	0.95	1.40	1.59	1.71	0.86	1.17	1.28	1.33	0.84	0.29	0.08	0.212
	SOS	4.39	9.42	12.89	16.01	4.30	9.19	12.68	15.78	2.29	1.48	0.90	0.78
	MBSS	2.75	7.05	9.61	12.85	2.68	6.86	9.39	12.59	2.27	1.43	0.89	0.76
	API-MBSS	**4.34**	**9.31**	**12.30**	**15.63**	**4.23**	**9.05**	**12.05**	**15.41**	**2.28**	**1.45**	**0.89**	**0.78**
Train	**BSS**	1.24	1.43	1.59	1.67	1.12	1.20	1.29	1.32	0.17	0.25	0.30	0.099
	SOS	5.54	8.04	11.91	15.74	5.34	7.81	11.73	15.53	2.15	1.41	1.13	0.89
	MBSS	4.07	5.73	9.55	11.78	3.88	5.53	9.35	11.59	2.15	1.40	1.34	0.86
	API-MBSS	**5.32**	**7.83**	**11.80**	**14.79**	**5.12**	**7.60**	**11.59**	**14.58**	**2.15**	**1.40**	**1.05**	**0.87**
Restaurant	**BSS**	2.07	1.59	1.84	1.64	1.75	1.27	1.37	1.30	0.48	.002	0.125	0.066
	SOS	1.72	7.20	9.81	15.13	1.48	7.00	9.59	14.95	1.42	1.02	0.82	2.79
	MBSS	2.66	6.02	9.54	11.18	2.52	5.84	9.29	10.90	1.11	1.27	0.65	2.70
	API-MBSS	**2.35**	**7.45**	**11.52**	**14.36**	**2.14**	**7.25**	**9.96**	**14.38**	**1.18**	**0.93**	**0.51**	**2.87**
Babble	**BSS**	1.47	1.52	1.63	1.74	1.19	1.21	1.29	1.35	0.35	0.42	0.073	0.023
	SOS	2.74	7.34	11.52	14.74	2.57	7.05	11.33	14.51	1.61	1.49	2.12	0.72
	MBSS	2.66	5.95	9.54	11.81	2.52	5.74	9.31	11.52	1.22	1.21	1.42	0.43
	API-MBSS	**3.06**	**7.45**	**11.52**	**14.36**	**2.87**	**7.17**	**11.31**	**14.11**	**1.36**	**1.44**	**1.35**	**0.71**
Airport	**BSS**	1.66	1.52	1.62	1.72	1.43	1.23	1.30	1.33	0.59	0.06	0.016	0.14
	SOS	3.60	8.30	11.04	15.68	3.33	8.06	10.85	15.48	1.80	0.91	1.06	0.75
	MBSS	3.93	6.75	9.06	12.04	3.78	6.53	8.81	11.77	1.28	0.90	1.05	0.77
	API-MBSS	**5.31**	**7.58**	**11.62**	**15.01**	**5.26**	**7.43**	**11.44**	**14.81**	**1.49**	**0.91**	**1.05**	**0.76**
Street	**BSS**	2.49	1.51	1.66	1.55	1.71	1.25	1.31	1.29	0.13	0.51	0.067	0.14
	SOS	0.02	7.14	11.17	14.61	-0.17	6.96	11.33	14.41	1.04	1.46	5.15	1.10
	MBSS	1.88	5.60	9.42	9.93	1.71	5.39	9.22	9.73	1.04	1.11	5.14	0.97
	API-MBSS	**9.81**	**7.19**	**11.50**	**12.21**	**9.81**	**7.02**	**11.36**	**12.0**	**1.04**	**1.24**	**5.15**	**1.01**
Exhibition	**BSS**	2.08	1.63	1.66	1.74	2.01	1.34	1.35	1.37	0.11	0.10	0.38	0.07
	SOS	1.29	7.31	11.13	15.12	1.07	7.10	10.9	14.91	1.69	1.10	1.32	0.60
	MBSS	2.24	7.18	9.09	12.36	2.05	6.99	8.92	12.13	1.68	1.06	1.21	0.51
	API-MBSS	**7.04**	**9.13**	**11.07**	**15.40**	**7.01**	**8.99**	**10.94**	**15.20**	**1.69**	**1.10**	**1.32**	**0.58**
White	**BSS**	1.42	1.59	1.70	1.78	1.18	1.31	1.34	1.38	0.60	0.38	0.25	0.214
	SOS	6.98	10.30	13.65	16.67	6.75	10.1	13.43	16.45	2.46	1.94	1.46	0.99
	MBSS	6.10	8.80	11.93	13.46	5.90	8.63	11.77	13.26	2.40	1.88	1.45	0.99
	API-MBSS	**4.28**	**10.26**	**13.61**	**16.81**	**4.07**	**10.0**	**13.40**	**16.63**	**2.47**	**1.95**	**1.43**	**1.10**

Table 5.5.

SNR Improvement (Global), and SegSNR IMPROVEMENT RESULTS of Enhanced Speech Signals at (0, 5, 10, 15) dB SNRs. *sp10.wav* Utterance Produced by a male Speaker is Used as Original Speech Signal.

Noise Type	**Enhancement Algorithms**	**SNR improvement (dB)**				**SegSNR improvement (dB)**			
		0 dB	**5 dB**	**10 dB**	**15 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**
Car	**BSS**	0.95	-3.6	-8.41	-13.29	0.86	-3.83	-8.72	-13.67
	SOS	4.39	4.42	2.89	1.01	4.30	4.19	2.68	0.78
	MBSS	2.75	2.05	-0.39	-2.15	2.68	1.86	-0.61	-2.41
	API-MBSS	**4.34**	**4.31**	**2.30**	**0.63**	**4.23**	**4.05**	**2.05**	**0.41**
Train	**BSS**	1.24	-3.57	-8.41	-13.33	1.12	-3.8	-8.71	-13.68
	SOS	5.54	3.04	1.91	.74	5.34	2.81	1.73	0.53
	MBSS	4.07	0.73	-0.45	-3.22	3.88	0.53	-0.65	-3.41
	API-MBSS	**5.32**	**2.83**	**1.80**	**-0.21**	**5.12**	**2.60**	**1.59**	**-0.42**
Restaurant	**BSS**	2.07	-3.41	-8.16	-13.36	1.75	-3.73	-8.63	-13.7
	SOS	1.72	2.20	0.19	0.13	1.48	2.00	-0.41	-0.05
	MBSS	2.66	1.02	-0.46	-3.82	2.52	0.84	-0.71	-4.1
	API-MBSS	**2.35**	**2.45**	**1.52**	**-0.64**	**2.14**	**2.25**	**-0.04**	**-0.62**
Babble	**BSS**	1.47	-3.48	-8.37	-13.26	1.19	-3.79	-8.71	-13.65
	SOS	2.74	2.34	1.52	-0.26	2.57	2.05	1.33	-0.49
	MBSS	2.66	0.95	-0.46	-3.19	2.52	0.74	-0.69	-3.48
	API-MBSS	**3.06**	**2.45**	**1.52**	**-0.64**	**2.87**	**2.17**	**1.31**	**-0.89**
Airport	**BSS**	1.66	-3.48	-8.38	-13.28	1.43	-3.77	-8.7	-13.67
	SOS	3.60	3.30	1.04	0.68	3.33	3.06	0.85	0.48
	MBSS	3.93	1.75	-0.94	-2.96	3.78	1.53	-1.19	-3.23
	API-MBSS	**5.31**	**2.58**	**1.62**	**0.01**	**5.26**	**2.43**	**1.44**	**-0.19**
Street	**BSS**	2.49	-3.49	-8.34	-13.45	1.71	-3.77	-8.69	-13.71
	SOS	0.02	2.14	1.17	-0.39	-0.17	1.96	1.33	-0.59
	MBSS	1.88	0.60	0.58	-5.07	1.71	0.39	-0.78	-5.27
	API-MBSS	**9.81**	**2.19**	**1.50**	**-2.79**	**9.81**	**2.02**	**1.36**	**-3**
Exhibition	**BSS**	2.08	-3.37	-8.38	-13.26	2.01	-3.66	-8.65	-13.63
	SOS	1.29	2.31	1.13	0.12	1.07	2.10	0.9	-0.09
	MBSS	2.24	2.18	-0.91	-2.64	2.05	1.99	-1.08	-2.87
	API-MBSS	**7.04**	**4.13**	**1.07**	**0.40**	**7.01**	**3.99**	**0.94**	**0.20**
White	**BSS**	1.42	-3.41	-8.3	-13.22	1.18	-3.69	-8.66	-13.62
	SOS	6.98	5.30	2.63	1.67	6.75	5.1	3.43	1.45
	MBSS	6.10	3.80	1.93	-1.54	5.90	3.63	1.77	-1.74
	API-MBSS	**4.28**	**5.26**	**3.61**	**1.81**	**4.07**	**5.00**	**3.40**	**1.63**

TABLE 5.6.

RESULTS OF NOISE REDUCTION SPEECH QUALITY TEST.

Noise Type	**Enhancement Algorithms**	**PESQ Score**				**MOS Score**			
		0 dB	**5 dB**	**10 dB**	**15 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**
Car	**BSS**	1.749	1.925	2.154	2.213				
	SOS	1.622	2.163	2.579	2.831	2.5	3.3	4	4.4
	MBSS	1.496	1.982	2.259	2.602	2.2	3	3.8	4.3
	API-MBSS	**1.435**	**1.625**	**2.432**	**2.640**	**1.1**	**2**	**3.7**	**4.5**
Train	**BSS**	1.873	1.666	2.079	2.156				
	SOS	1.720	1.888	2.394	2.730	2.3	3.2	4	4.5
	MBSS	1.513	1.696	2.129	2.382	2.2	3	3.9	4.3
	API-MBSS	**1.522**	**1.611**	**2.143**	**2.327**	**1.6**	**2.5**	**3.6**	**4.3**
Restaurant	**BSS**	1.682	1.843	2.002	2.165				
	SOS	1.785	2.157	2.362	2.811	2.6	3.5	4	4.4
	MBSS	1.842	2.062	2.367	2.603	2.4	3.2	4	4.3
	API-MBSS	**1.633**	**1.958**	**2.321**	**2.563**	**2.6**	**3.3**	**4.1**	**4.5**
Babble	**BSS**	1.481	1.924	2.110	2.215				
	SOS	1.903	2.209	2.562	2.699	2	3	3.9	4.4
	MBSS	1.812	2.208	2.394	2.650	1.8	2.8	3.9	4.4
	API-MBSS	**1.334**	**2.105**	**2.436**	**2.657**	**1**	**2.2**	**3.6**	**4.3**
Airport	**BSS**	1.407	1.939	2.092	2.204				
	SOS	1.891	2.222	2.476	2.836	2.6	3.4	4	4.2
	MBSS	1.790	2.106	2.323	2.681	1.7	2.9	3.6	4.4
	API-MBSS	**1.462**	**2.117**	**2.424**	**2.729**	**2**	**3.3**	**3.9**	**4.5**
Street	**BSS**	1.511	1.833	2.045	2.018				
	SOS	1.578	2.013	2.406	2.536	2.4	2.9	3.8	4.4
	MBSS	1.592	1.933	2.249	2.213	1.9	2.8	3.5	4.3
	API-MBSS	**1.402**	**2.017**	**2.304**	**2.297**	**2.2**	**3.1**	**3.9**	**4.4**
Exhibition	**BSS**	1.721	1.655	2.109	2.127				
	SOS	1.799	2.004	2.267	2.694	2.4	3.2	3.8	4.2
	MBSS	1.527	1.977	1.968	2.517	2	2.9	3.8	4.4
	API-MBSS	**1.488**	**1.979**	**2.097**	**2.716**	**2**	**3.1**	**3.8**	**4.3**
White	**BSS**	1.663	1.957	2.087	2.151				
	SOS	1.912	2.232	2.483	2.800	3.0	3.7	4.2	4.4
	MBSS	1.655	1.971	2.303	2.563	3	3.7	4.3	4.5
	API-MBSS	**1.535**	**1.825**	**2.229**	**2.676**	**2.2**	**3.3**	**3.9**	**4.4**

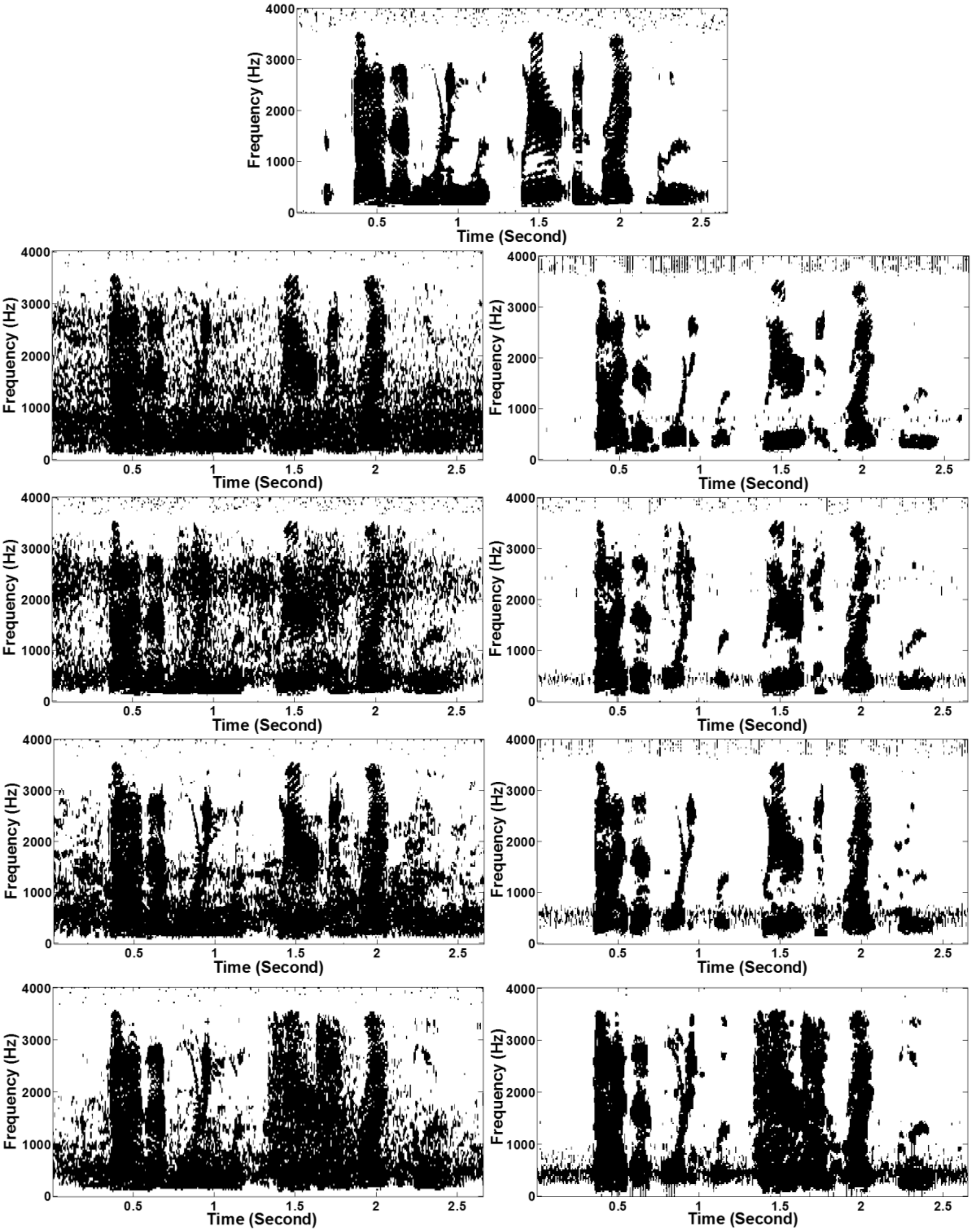


Fig. 5.8 (I): Speech spectrogram of *sp10.wav* utterance, "The sky that morning was clear and bright blue", by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech; (b, d, f, h) speech degraded by car noise, train noise, babble noise and restaurant noise, respectively (10 dB SNR); and (c, e, g, i) corresponding enhanced speech.

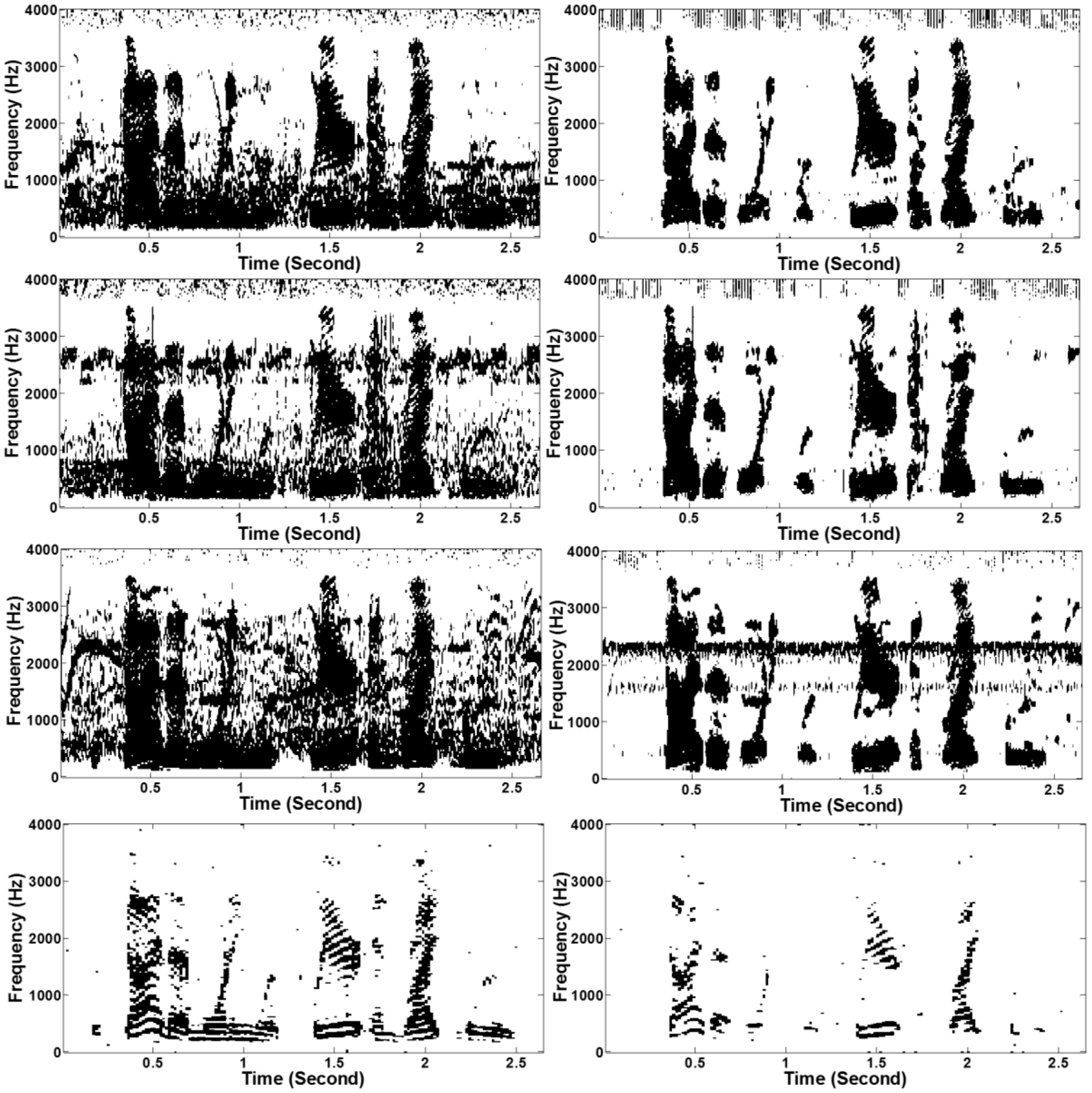


Fig. 5.8 (II): Speech spectrogram of *sp10.wav* utterance, "The sky that morning was clear and bright blue", by a male speaker from the NOIZEUS corpus (From top to bottom): (j, l, n, p) speech degraded by airport noise, street noise, exhibition noise, and white noise respectively (10 dB SNR); and (k, m, o, q) corresponding enhanced speech.

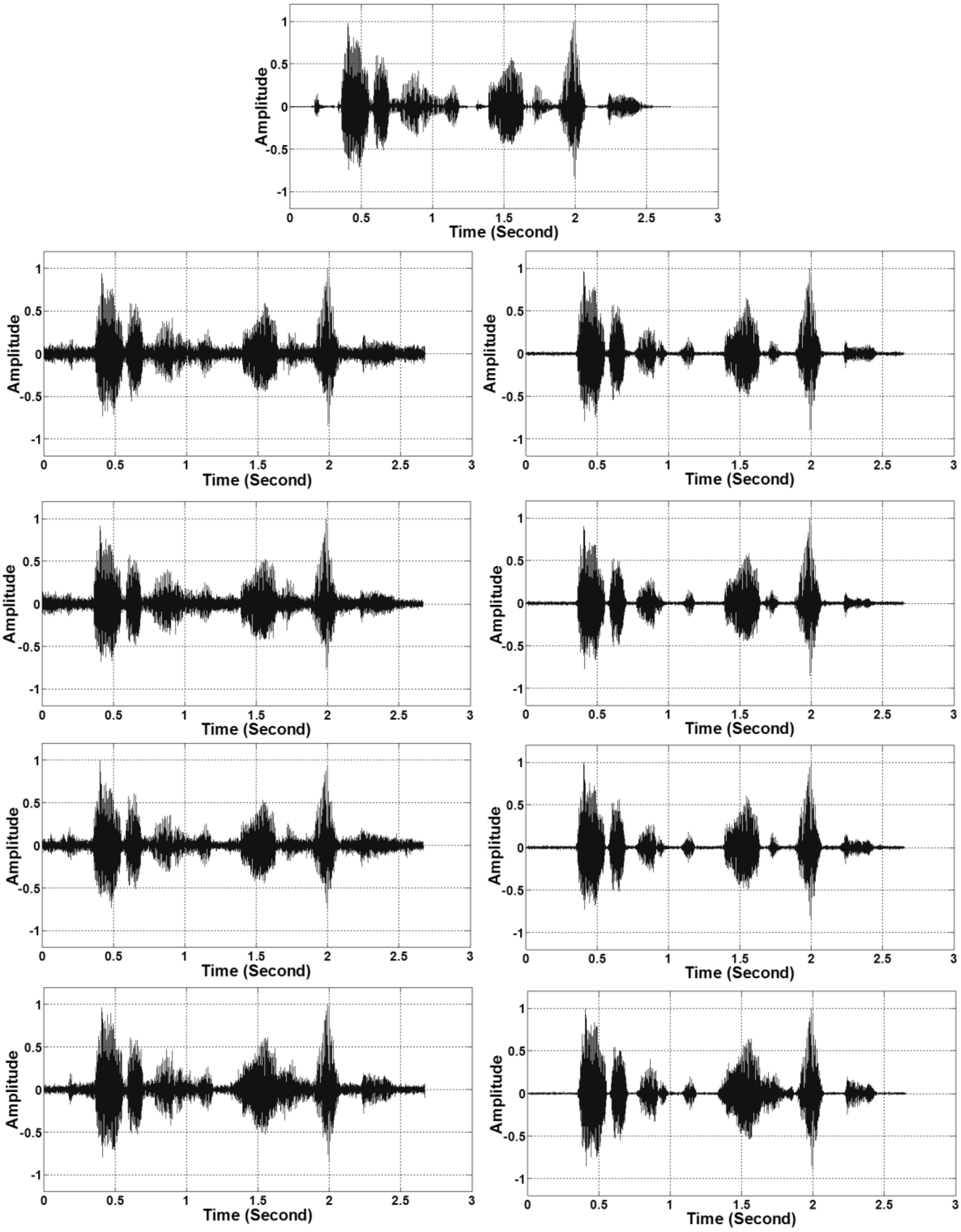


Fig. 5.9 (I): Temporal waveforms of *sp10.wav* utterances, "The sky that morning was clear and bright blue", by a male speaker from the NOIZEUS corpus (From top to bottom): (b, d, f, h) speech degraded by car noise, train noise, babble noise, and restaurant noise, respectively (10 dB SNR); and (c, e, g, i) corresponding enhanced speech.

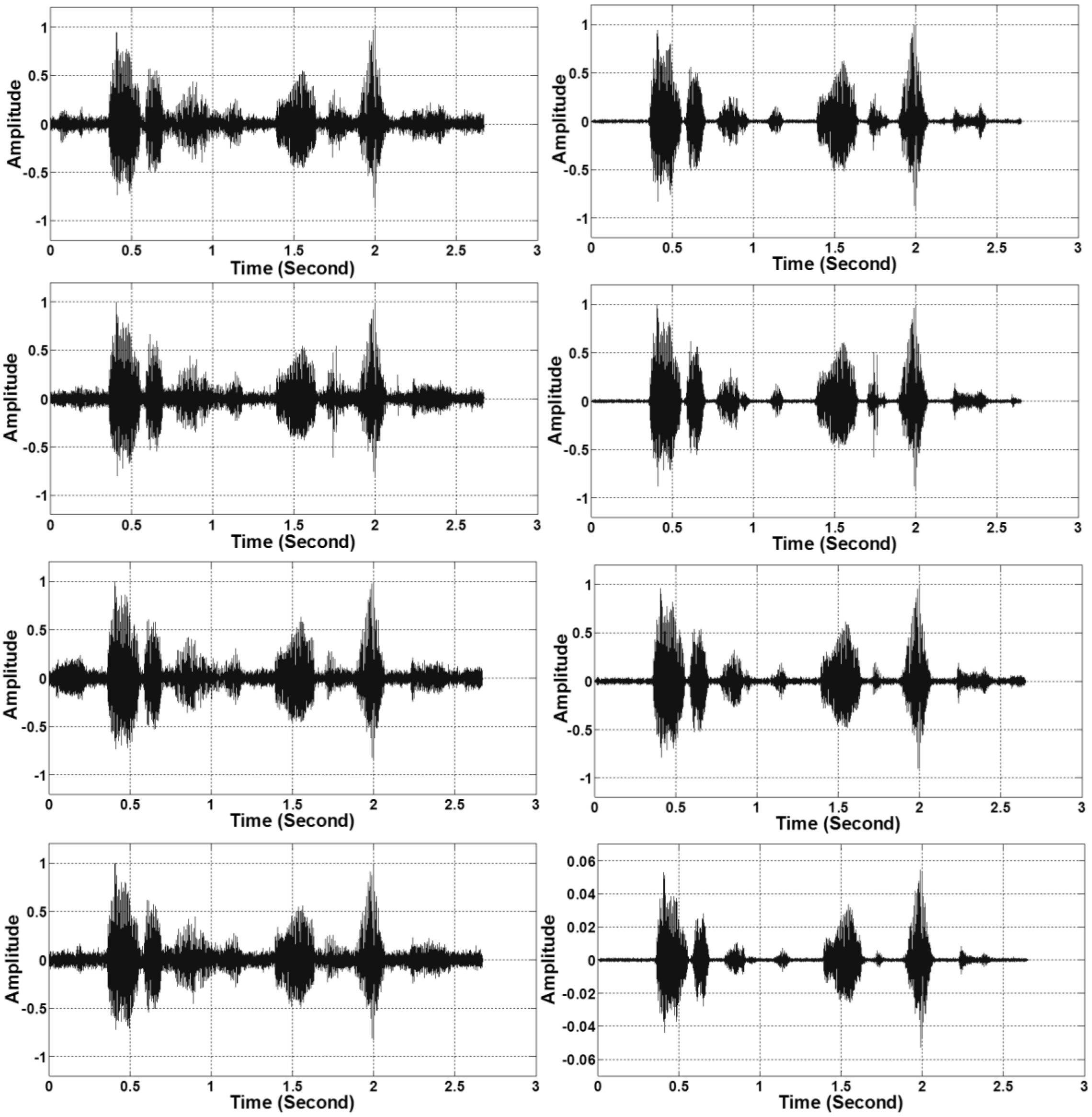


Fig. 5.9 (II): Temporal waveforms of *sp10.wav* utterances, "The sky that morning was clear and bright blue", by a male speaker from the NOIZEUS corpus (From top to bottom): (j, l, n, p) speech degraded by airport noise, street noise, exhibition noise, and white noise respectively (10 dB SNR); and (k, m, o, q) corresponding enhanced speech.

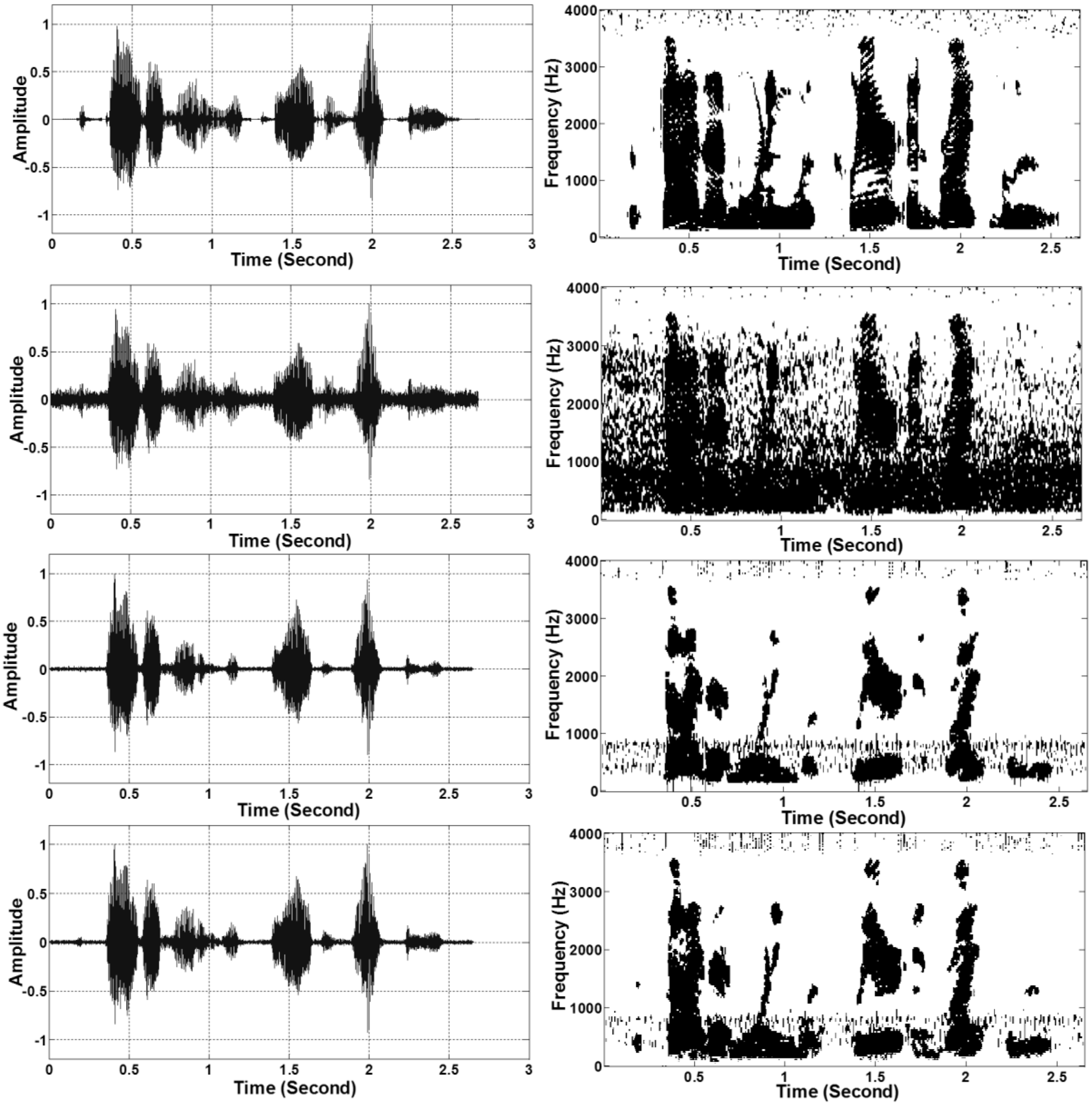


Fig. 5.10 (I): Temporal waveforms and speech spectrogram of *sp10.wav* utterance, "The sky that morning was clear and bright blue", by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech (PESQ = 4.5); (b) speech degraded by car noise (10 dB SNR) (PESQ = 2.169); (c) speech enhanced by MBSS (PESQ = 2.259); (d) speech enhanced by MBSS (with averaging based noise estimation replaced by adaptive noise estimation).

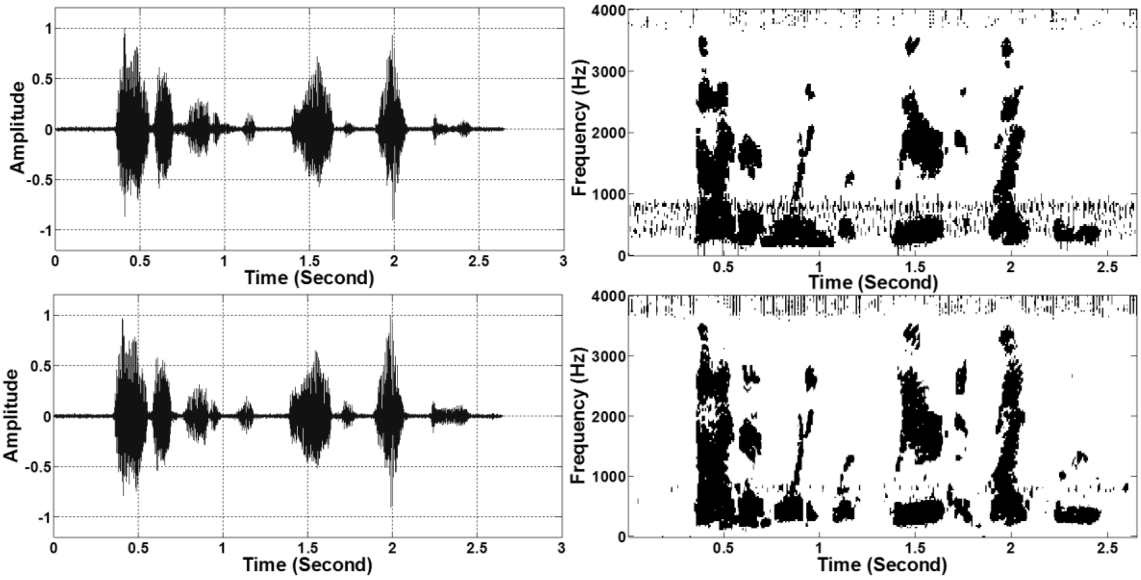


Fig. 5.10 (II): Temporal waveforms and speech spectrogram of *sp10.wav* utterance, "The sky that morning was clear and bright blue", by a male speaker from the NOIZEUS corpus (From top to bottom): (e) speech enhanced by API-MBSS (with adaptive noise estimation replaced by averaging approach); and (f) speech enhanced by API-MBSS (PESQ = 2.432**)**.

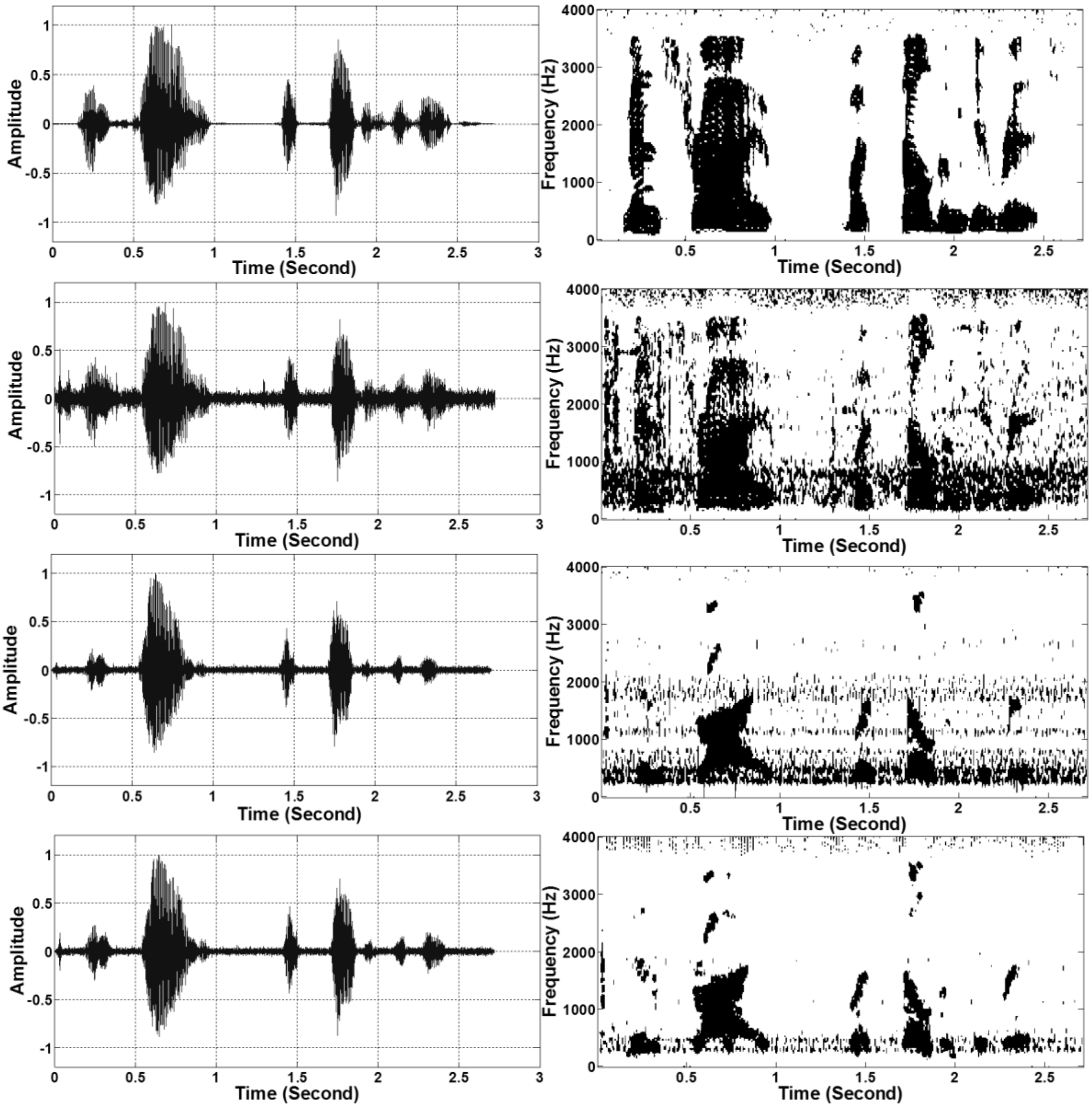


Fig. 5.11 (I): Temporal waveforms and speech spectrogram with *sp 6.wav* utterance, "Men strive but seldom get rich", by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech (PESQ = 4.5); (b) noisy speech (degraded by car noise at 10 dB SNR) (PESQ = 2.205); (c) speech enhanced by MBSS (PESQ = 2.157); (d) speech enhanced by MBSS (with averaging based noise estimation replaced by adaptive noise estimation).

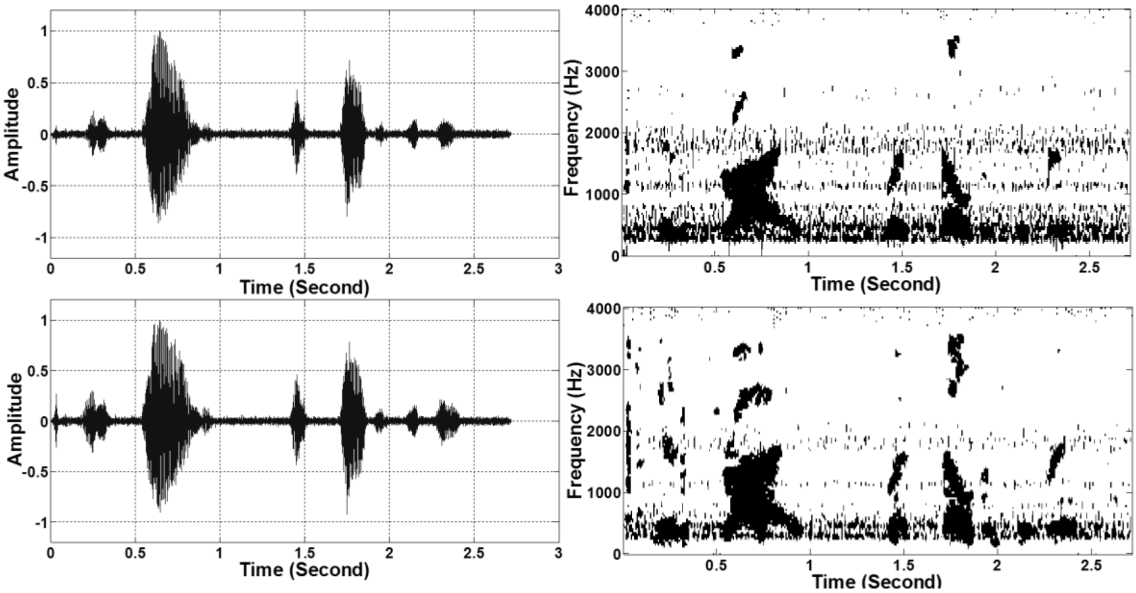


Fig. 5.11(II): Temporal waveforms and speech spectrogram with *sp6.wav* utterance, "Men strive but seldom get rich", by a male speaker from the NOIZEUS corpus (From top to bottom): (e) speech enhanced by API-MBSS (with adaptive noise estimation replaced by averaging approach); and (f) speech enhanced by API-MBSS (PESQ = 2.330).

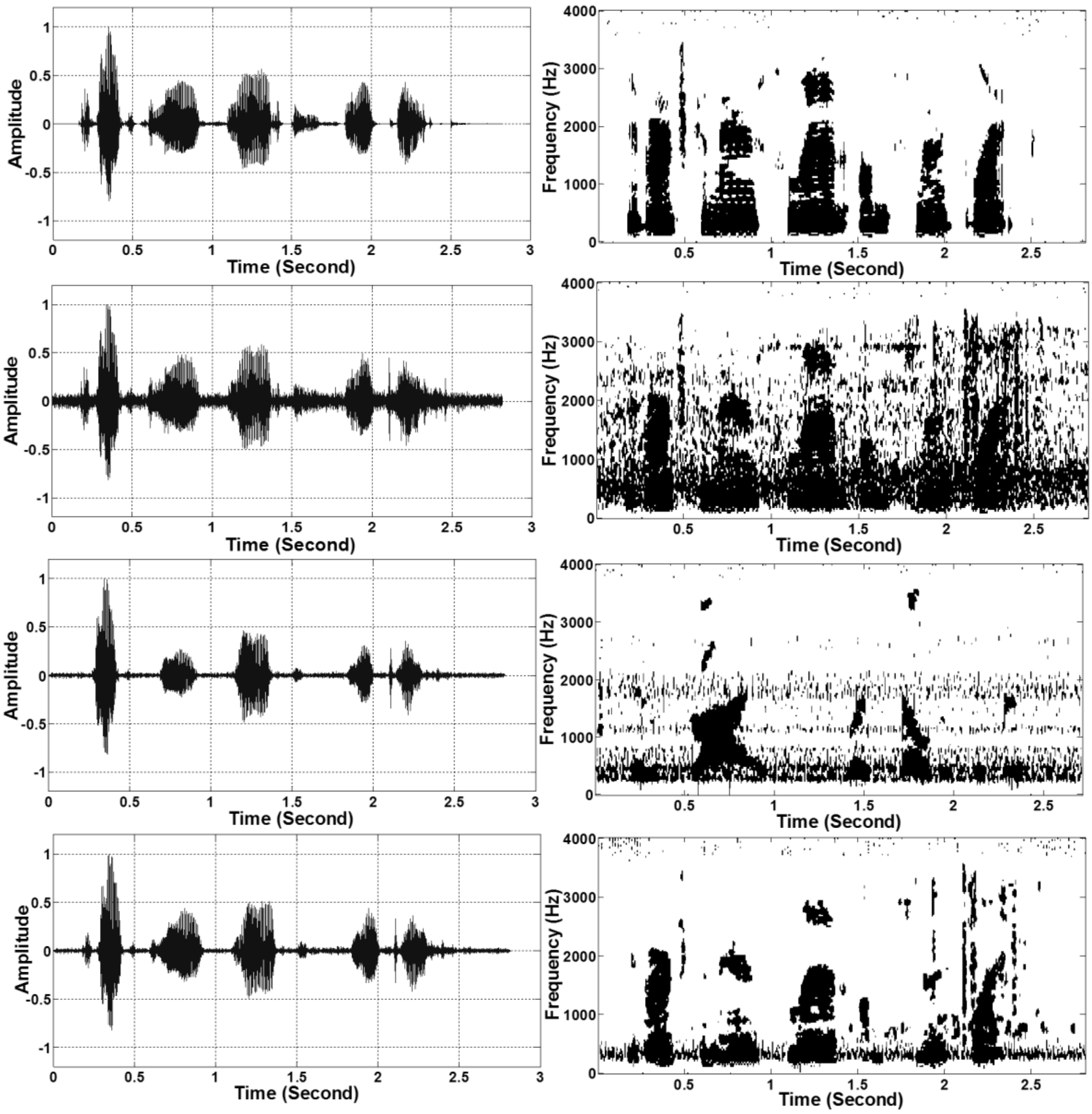


Fig. 5.12 (I): Temporal wave forms and speech spectrogram of *sp1.wav* utterance, "The birch canoe slid on the smooth planks", by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech (PESQ = 4.5); (b) noisy speech (degraded by car noise at 5 dB SNR); (c) speech enhanced by MBSS (PESQ = 2.030); (d) speech enhanced by MBSS (with averaging based noise estimation replaced by adaptive noise estimation).

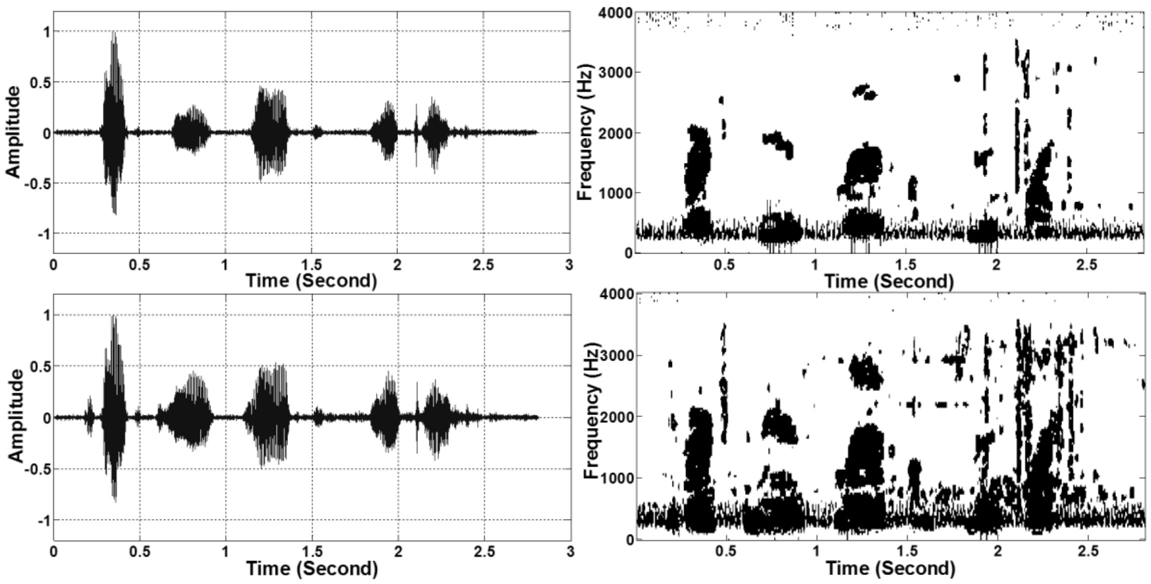


Fig. 5.12 (II): Temporal waveforms and speech spectrogram of *sp1.wav* utterance, "The birch canoe slid on the smooth planks", by a male speaker from the NOIZEUS corpus (From top to bottom): (e) speech enhanced by API-MBSS (with adaptive noise estimation replaced by averaging approach); and (f) speech enhanced by API-MBSS (PESQ = 2.169).

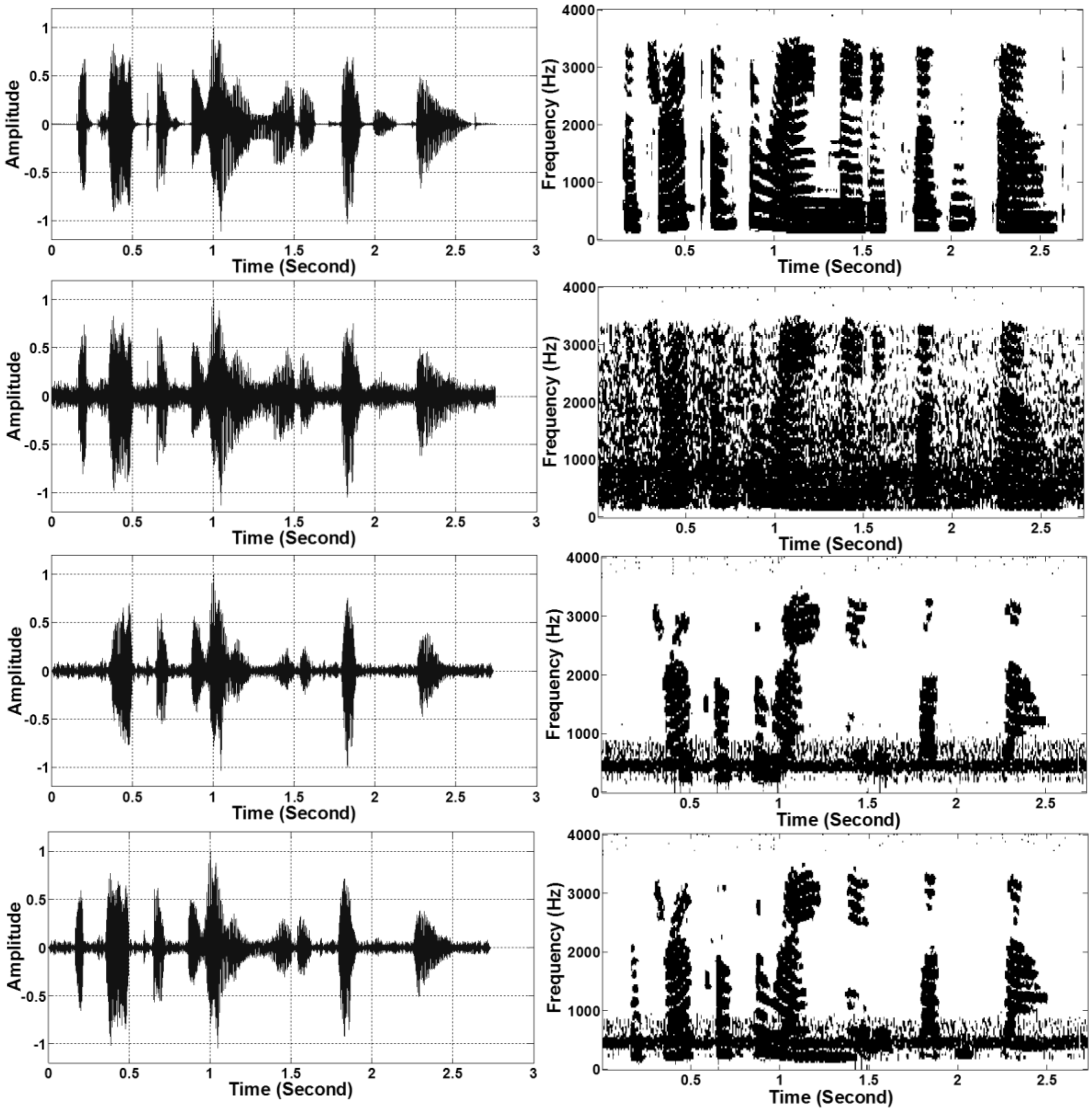


Fig. 5.13 (I): Temporal waveforms and speech spectrogram of *sp12.wav* utterance, "The drip of the rain made a pleasant sound", by a female speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech (PESQ = 4.5); (b) noisy speech (degraded by car noise at 10 dB SNR) (PESQ = 2.043); (c) speech enhanced by MBSS (PESQ = 2.005); (d) speech enhanced by MBSS (with averaging based noise estimation replaced by adaptive noise estimation).

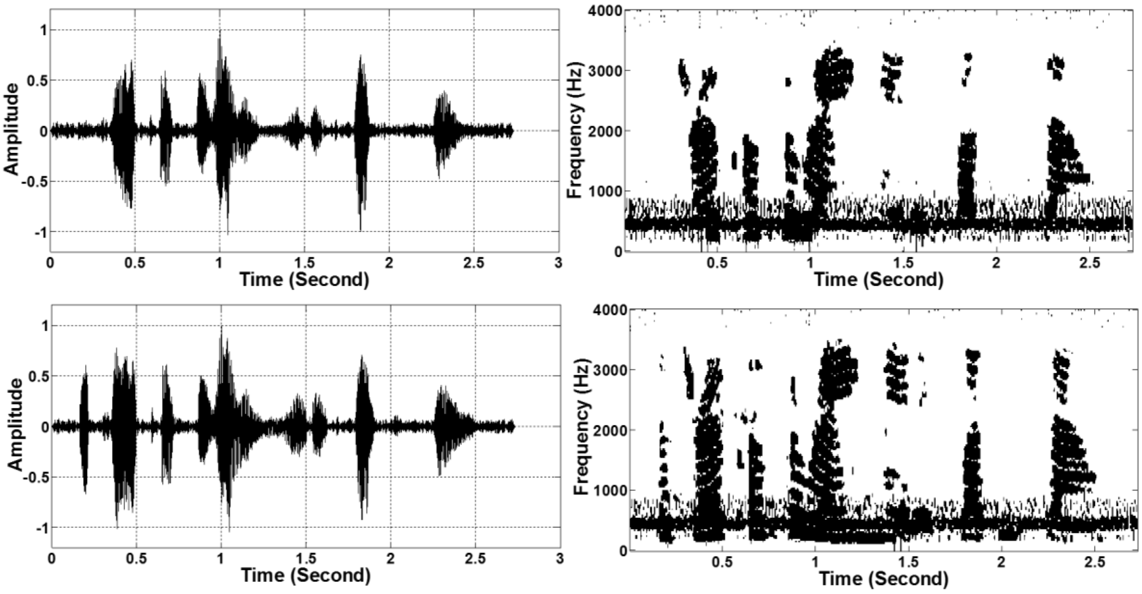


Fig. 5.13 (II): Temporal waveforms and speech spectrogram of *sp12.wav* utterance, "The drip of the rain made a pleasant sound", by a female speaker from the NOIZEUS corpus (From top to bottom): (e) speech enhanced by API-MBSS (with adaptive noise estimation replaced by averaging approach); and (f) speech enhanced by API-MBSS (PESQ = 2.483).

5.5. Summary

In this chapter, critical band rate scale (an auditory perception) based improved multi-band spectral subtraction, API-MBSS, algorithm is presented for enhancement of speech degraded by non-stationary or colored noise. As the sensitivity of human ear is a non-linear function of frequency, the bands in API-MBSS has been kept non-uniformly frequency spaced. The proposed algorithm, API-MBSS, reduces noise tones efficiently that appear in the case of uniformly frequency spaced multi-band spectral subtraction (MBSS), BSS and SOS algorithms.

In our algorithm, the three adjacent critical bands are grouped together into six non-uniformly spaced frequency bands that broadly resembles with the non-uniform frequency spacing given by the human auditory system. The algorithm presented in this chapter uses a noise estimation algorithm that does not the need of voice activity detection for noise estimation. This approach estimates and updates the noise spectrum continuously, even during speech activity from each band.

Simulations with different type of noises and MOS scores of the subjective listening test reveal that the API-MBSS algorithm reduces the remnant noise tones efficiently that appear in the case of conventional multi-band spectral subtraction algorithm, BSS, SOS, and improves the overall quality of degraded speech at low SNRs. Furthermore, the API-MBSS has strong flexibility to adapt any complicated rigorous speech environment by adjusting the over-subtraction factor for each non-uniformly frequency spaced band, separately.

The API-MBSS shows superior results for all types noises (non-stationary and stationary) at low SNR (≤ 10 dB) except restaurant noise, while for SNR (≈ 15 dB) it shows better results for car, exhibition, white (stationary) noises and gives poorer results for street noise. Also, this proposed algorithms shows better performance than multi-band spectral subtraction and basic spectral subtraction algorithm at all levels of SNRs; also, gives comparable performance with spectral over-subtraction algorithm except for street noise case.

In the next chapter, a perceptually motivated stationary WPT based improved multi-band spectral over-subtraction algorithm is proposed for enhancement of degraded speech. The stationary WPT and its advantages over WPT have already been described in Chapter 3.

Chapter 6

Perceptually Motivated Stationary WPT based Improved Spectral Over-Subtraction

6.1. Introduction

The front-end time-frequency decomposition for the algorithms proposed in chapters 4 and 5 is obtained using STFT, and the speech enhancement is performed on the transform coefficients. For analyzing non-stationary signals with frequent transients, such as speech, wavelet transform have been proved to be is an important tool and has been used in various speech applications [30-36]. The wavelet transform divides a signal into different frequency components, and each component can be analyzed with a resolution matched to its scale. Since the speech signal is non-stationary in nature and contain transients, therefore spectral subtraction algorithms with wavelet transform based time-frequency decomposition is likely to give better performance than classical methods employing STFT. This is more so, as the background noise taken for our speech enhancement algorithms is also of non-stationary nature.

Because of ease of implementation using binary filterbank structures, discrete wavelet transform (DWT) has been used for many speech applications structures. It also has the advantage

The work reported in this chapter has resulted in the following publication

[I]. Navneet Upadhyay, and Abhijit Karmakar, "A perceptually motivated stationary wavelet packet filterbank utilizing improved spectral over-subtraction algorithm for enhancing speech in non-stationary environments," in *Proceedings of 4th IEEE International Conference on Intelligent Human Computer Interaction*, IIT Kharagpur, India, Dec. 27─29, 2012, pp. 392─398.

in the context of auditory inspired time-frequency decomposition because of its octave band or logarithmic spectrum. In this context, the extension of DWT, namely wavelet packet transform (WPT), is suitable for matching the auditory frequency scale more closely with the frequency bands available in WPT. The over-sampled filterbank realization of WPT i.e., stationary WPT (SWPT) has been utilized instead of the critically sampled WPT. The advantage of using SWPT is that it does not suffer from the *shift-variance* problem as in WPT. Also, both for the low and high frequency band, we can work with maximum temporal resolution available. The stationary wavelet packet transform (SWPT) overcomes the *shift-invariance* problem by removing the down-sampling at each decomposition level [63, 64]. Also, the length of the approximation and detail coefficients at each level are of same length as the incoming signal. In general, the major drawback of speech enhancement in wavelet packet domain is that it contains substantial signal distortion because of down-sampling at each level of decomposition and high computational load, as described in Chapter 3. Although perceptual wavelet packet decomposition (PWPD) leads to improved speech quality and reduces computational load but speech distortion causes loss of information, caused by down-sampling, still remains as a problem which reduces the intelligibility and perceptual quality of enhanced speech signal [30, 32, 34].

It is well-known that the human auditory system operates like a non-uniform filterbank and humans are capable of detecting the desired speech in various noisy environments. This has been exploited in the proposed algorithm in chapter 5. In this chapter also, the perceptual properties of human auditory system has been exploited and the SWPT is obtained to closely match the perceptual frequency scale. The proposed algorithm in this chapter also employs the improved spectral over-subtraction algorithm to estimate the noise from each subband by using the noise estimate adaptively. The proposed algorithm, thus, utilizes the characteristics of the subband signal in the temporal domain separately, in accordance to the perceptual frequency scale.

In [67], a speech enhancement system is proposed which integrates bark scaled filterbank and Wiener filtering, and modified according to speech presence uncertainty. Bark scaled filterbank is obtained from a PWPD tree and adjusting five levels of full WPD tree of speech signal according to critical bands. The resultant decomposition tree structure is called as bark scaled wavelet packet decomposition (BS-WPD). The drawback of the Wiener filtering which is used for enhancement of frequency bands of bark scaled filterbank in [67] contains large remnant musical noise.

In [68], a method using un-decimated wavelet packet perceptual filterbanks and minimum mean square error short-time spectral amplitude (MMSE-STSA) estimation is proposed to enhance the degraded speech. The drawback of this method is that it uses the voice activity detector (VAD) for estimation of noise power spectral density during speech pauses. The proposed perceptually motivated stationary WPT based speech enhancement algorithm does not require voice activity detection, and explained further below.

The perceptually motivated stationary WPT based improved multiband spectral over-subtraction (PMS-MBSS) algorithm proposed in this chapter where the front-end decomposition of the degraded speech in different subbands is performed by temporally processing it using a perceptually motivated stationary wavelet packet filterbank (PM-SWPFB). The proposed algorithm also incorporates an improved version of the spectral over-subtraction (I-SOS) algorithm for the reduction of the non-stationary or colored noise in the degraded speech. The PM-SWPFB is obtained by selecting the stationary wavelet packet tree in such a manner that it matches closely with the critical band structure of the psychoacoustic model of human auditory system. After the decomposition of the input noisy speech signal by the PM-SWPFB, the I-SOS algorithm is used to estimate clean speech from each of the subbands. The noise estimation technique for the I-SOS is done in each subband separately, without the need for explicit speech silence detection. Further, the subband noise estimate is updated by adaptively smoothing the noisy signal power. The smoothing parameter in each subband is controlled by a function of the estimated signal-to-noise ratio.

The performance of the proposed speech enhancement algorithm is evaluated objectively by signal-to-noise ratio and subjectively by subjective listening test. The results confirm that the proposed speech enhancement system is capable of reducing noise with small amount of speech degradation which remains acceptable in real-world environments. The overall performance of the proposed algorithm is found to be superior to several competitive methods, for some type of noise conditions.

The remaining part of the chapter is organized as follows. In Section 6.2, the proposed speech enhancement algorithm is described in detail. Section 6.3 explains the design of perceptually motivated stationary wavelet packet filterbank. In Section 6.4, the noise estimation approach is details. Section 6.5 describes the improved spectral over-subtraction algorithm in detail. In Section 6.6 experimental results and performance evaluation is given. In this section, sub-section (i.e. Section 6.5.1) gives the approach for selection of wavelet filter. Finally, Section 6.7 concludes this chapter.

6.2. Stationary WPT based Improved Spectral Over-Subtraction Speech Enhancement Algorithm

The block diagram of the perceptually motivated stationary WPT based improved multiband spectral over-subtraction (PMS-MBSS) speech enhancement algorithm, as proposed in this chapter is shown in Fig. 6.1. The steps of PMS-MBSS algorithm are as follows:

i) Firstly, the perceptually motivated stationary wavelet packet filterbank (PM-SWPFB) is applied to decompose the input noisy speech signal $y(n)$ into non-uniform subband signals $y_i(n)$. The detail of stationary WPT is explained in Section 3.7 of Chapter 3 and the construction of the perceptually motivated stationary (PM-SWPFB) has been explained in detail in Section 6.3 of this chapter.

ii) Secondly, we use an improved spectral over-subtraction (I-SOS) algorithm, to estimate the speech in each subband signal by using an adaptive noise estimation approach. The noise estimation approach is explained in section 6.4 and subsequently in section 6.5; the

I-SOS is presented. The block diagram of the speech enhancement in each subband is shown in Fig. 6.2.

iii) Finally, the enhanced speech signal $\hat{s}\,(n)$ is reconstructed by the stationary wavelet packet filterbank synthesis stage.

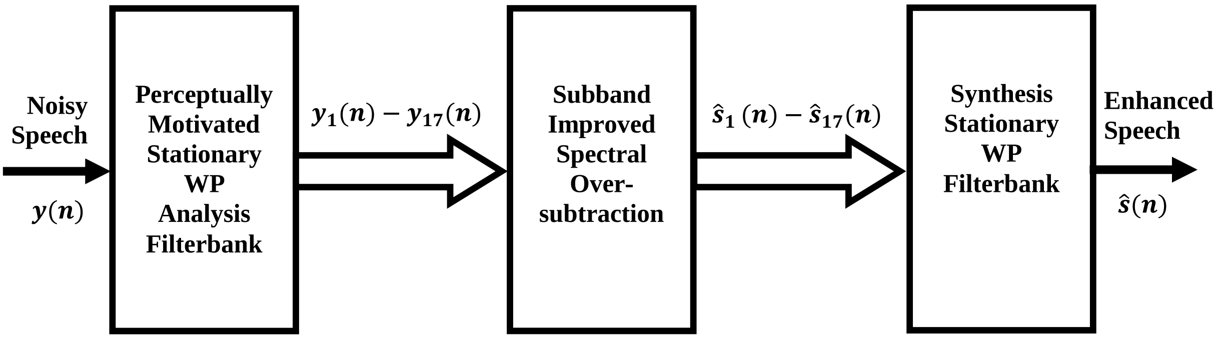

Noisy Speech
$y(n)$
Perceptually Motivated Stationary WP Analysis Filterbank
$y_1(n) - y_{17}(n)$
Subband Improved Spectral Over-subtraction
$\hat{s}_1\,(n) - \hat{s}_{17}(n)$
Synthesis Stationary WP Filterbank
Enhanced Speech
$\hat{s}(n)$

Fig. 6.1: Block diagram of proposed speech enhancement algorithm, PMS-MBSS.

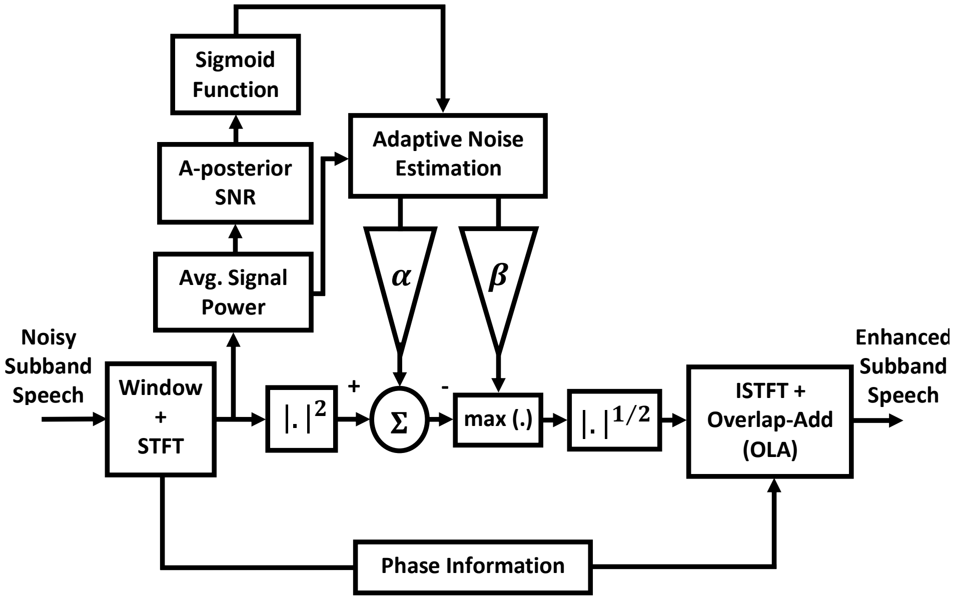


Fig. 6.2: Block diagram improved spectral over-subtraction algorithm.

6.3. Construction of Perceptually Motivated Stationary Wavelet Packet Filterbank (PM-SWPFB)

The unique feature of auditory processing of human is existence of critical bands (CBs). When the critical bands are placed next to each other, the critical band rate scale is obtained. This critical band rate scale is based on the fact that our hearing system analyses a broad spectrum in parts corresponding to CBs. Thus, for wavelet based speech processing, the WP tree is often chosen so that the incoming signal is analyzed and processed with bandwidths of one CB. Here, for the stationary WPT based decomposition, the wavelet packet tree is selected to approximate the CBs of the psychoacoustic model as close as possible, so that the signal can be analyzed and processed in accordance to the perceptual frequency scale.

The critical band scale is also known as Bark scale, where one Bark is referred as the bandwidth of one critical band. Based on the measurements by Zwicker *et* al. [84, 85], an approximate analytical expression to describe the relationship between linear frequency and critical band number (in Bark) is

$$z(f) = 13\arctan\left(\frac{0.76\times f}{1000}\right) + 3.5\arctan\left(\frac{f}{7500}\right)^2 \qquad (6.1)$$

Here, z is the critical band number in Bark, and f is the linear frequency in Hertz. The corresponding critical bandwidth (CBW), refers to the non-uniform frequency response of the human ear, of the center frequencies can be expressed by

$$\mathrm{CBW}(f_c) = 25 + 75\left(1 + 1.4\left(\frac{f_c}{1000}\right)^2\right)^{0.69} \qquad (6.2)$$

where f_c is the center frequency in Hertz. Theoretically, the range of audio frequency spreads from 20 Hz to 20 kHz. However, the frequency range of the narrowband speech of human voice is typically only from about 300 Hz to 3.4 kHz, where the sampling rate is chosen to be 8 kHz. Within the signal bandwidth, there are approximately 17 CBs as listed in Table 6.1 [85].

For a given sampling rate of f_s, the frequency bandwidth of stationary wavelet packet decomposition (SWPD) at the j^{th} level [65] is

$$\mathrm{CBW}(j) = \frac{f_s}{2^{j+1}} \tag{6.3}$$

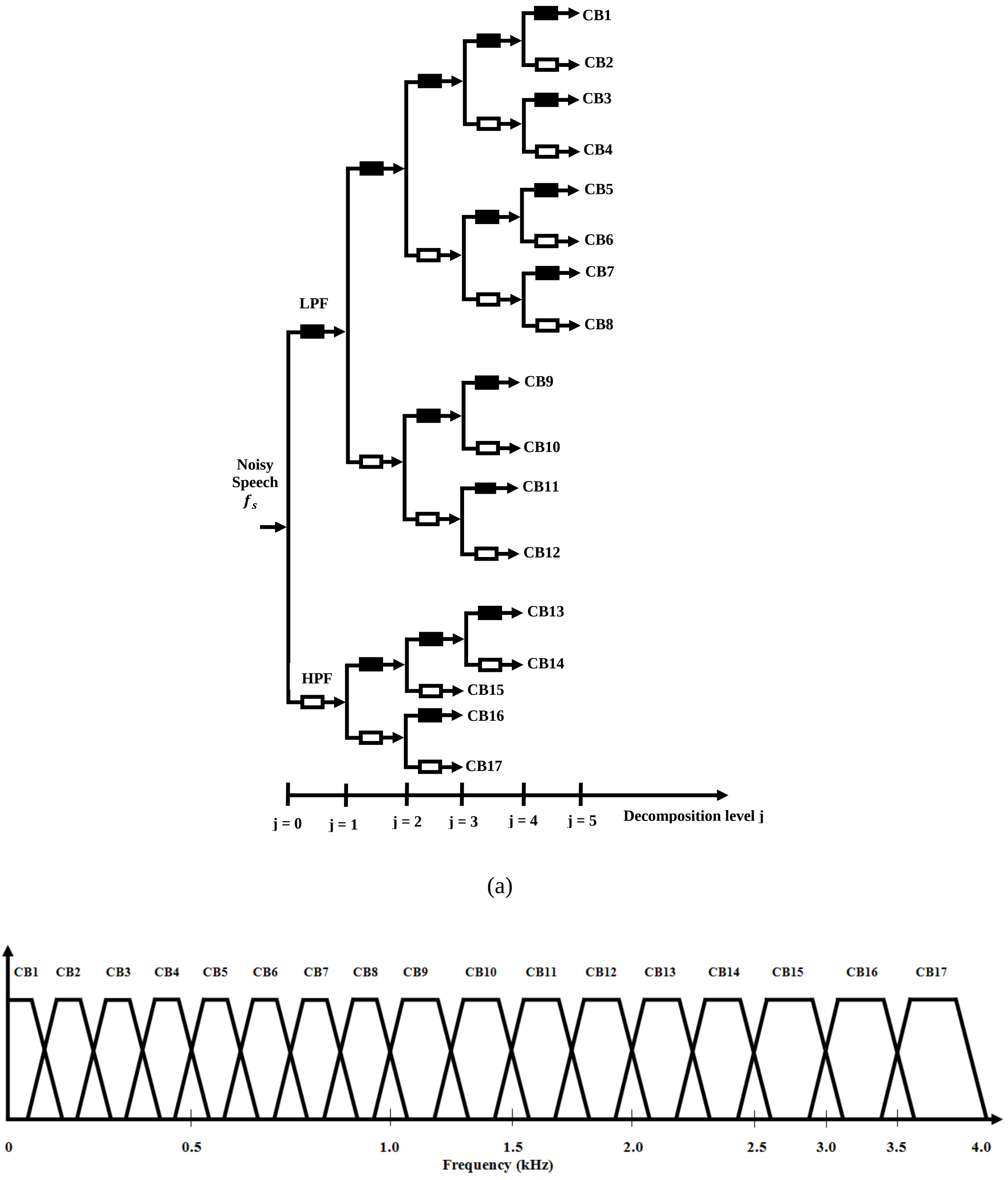
where $\mathrm{CBW}(j)$ represents the frequency bandwidth corresponding to level and node (j, l) in SWPT tree, J is the maximum number of levels $j = 0,1,2, \dots, J$; l is the position of node or shift, where $l = 0,1, \dots, (2^j - 1)$, and f_s is the sampling rate.

According to the specifications of center frequencies (f_c), CBW (Δf), lower (f_l), and upper (f_u) cut-off frequencies given in Table 6.1, the tree structure of the PM-SWPFB can be constructed as shown in Fig. 6.3 (a). The corresponding frequency bandwidth of the PM-SWPFB is shown in Fig. 6.3 (b).

TABLE 6.1.

CB RATE SCALE, z AND PERCEPTUALLY MOTIVATED STATIONARY WAVELET PACKET TREE FOR SAMPLING FREQUENCY OF 8 KHZ.

z [Bark]	Zwicker's Critical Band (Hz)				Perceptually Motivated Wavelet Packet Tree (Hz)			
	f_l	f_u	f_c	Δf	f_l	f_u	f_c	Δf
1	0	100	50	100	0	125	62.5	125
2	100	200	150	100	125	250	187.5	125
3	200	300	250	100	250	375	312.5	125
4	300	400	350	100	375	500	437.5	125
5	400	510	450	110	500	625	562.5	125
6	510	630	570	120	625	750	687.5	125
7	630	770	700	140	750	875	812.5	125
8	770	920	840	150	875	1000	937.5	125
9	920	1080	1000	160	1000	1250	1125	250
10	1080	1270	1170	190	1250	1500	1375	250
11	1270	1480	1370	210	1500	1750	1625	250
12	1480	1720	1600	240	1750	2000	1875	250
13	1720	2000	1850	280	2000	2250	2125	250
14	2000	2320	2150	320	2250	2500	2375	250
15	2320	2700	2500	380	2500	3000	2750	500
16	2700	3150	2900	450	3000	3500	3250	500
17	3150	3700	3400	550	3500	4000	3750	500



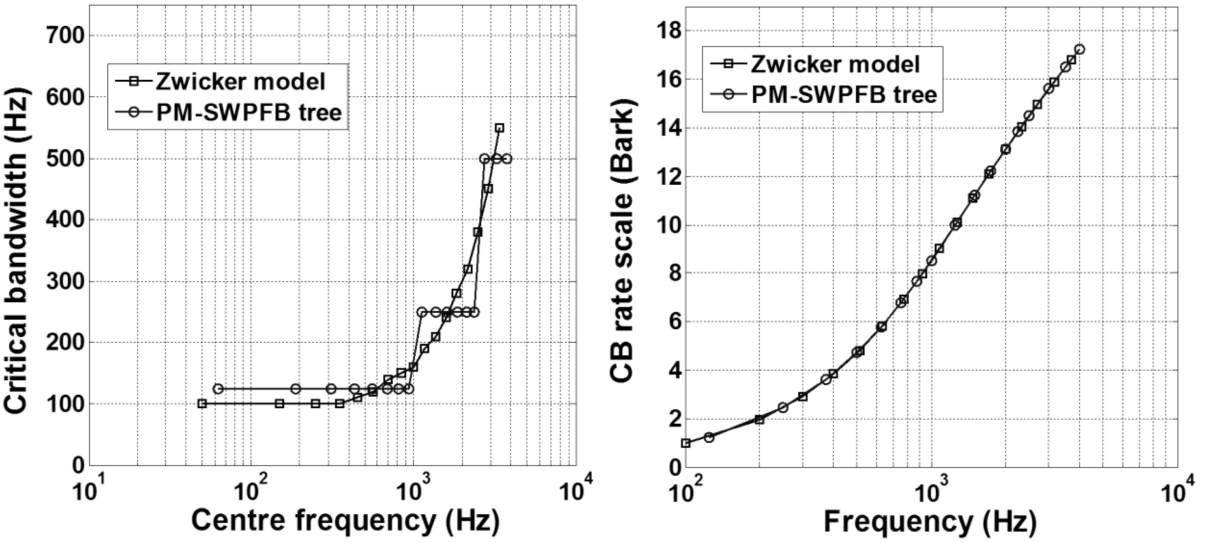
(a)

(b)

Fig. 6.3: (a) The tree structure of the proposed PM-SWPFB, and (b) the frequency bandwidths for the PM-SWPFB tree.

It contains 5 decomposition stages to approximate these 17 CBs which correspond to wavelet packet coefficient sets $w_{j,m}\ (k)$. More precisely, $w_{j,m}\ (k)$ defines the k^{th} coefficient of the m^{th} subband at j^{th} decomposition stage of PM-SWPFB, where $j = 3, 4, 5$, $m = 1,2,\ldots,17$, and $k = 1,\ldots,N/2^{j}$.

The resulting 17 band PM-SWPFB of the critical band rate scale and the CBW are plotted in Fig. 6.4 (a) and Fig. 6.4 (b), respectively. It can be observed easily observed that the perceptually motivated wavelet tree closely matches the critical band structure as proposed by Zwicker [85].



(a) (b)

Fig. 6.4: Comparison of the perceptually motivated stationary wavelet packet filterbank decomposition tree with Zwicker's model of perceptual frequency scale: (a) critical bandwidth as a function of centre frequency, and (b) critical band rate scale as a function of frequency.

6.4. Noise Estimation

The noise estimate can have a major impact on the quality of enhanced signal. If the noise estimate is too low, annoying remnant noise will be audible. Whereas, if the noise estimate is too high, speech will be distorted, and possibly would result in intelligibility loss [3]. The simplest approach is to estimate and update the noise spectrum during the silence segments of the signal using voice activity detection (VAD) algorithm. But the drawback of this approach is that it works satisfactorily only in case of stationary noise, and does not work well in more realistic environments (non-stationary noise), as discussed in Section 2.4.

In the traditional approach, the noise is estimated with recursive averaging, where the noise spectrum is estimated as a weighted average of the past noise estimates and the current noisy speech spectrum. The weights change adaptively depending on the effective *a-posteriori* SNR of each frequency bin or the probability of speech presence. In our proposed algorithm, PMS-MBSS, we have estimated and updated the noise spectrum in each frequency subband, separately.

The noise estimation as in (5.1) has been modified as below, for each subband using the first order relation as

$$|\widehat{D}_i(\omega,k)|^2 = \lambda_i(\omega,k)|\widehat{D}_i(\omega,k-1)|^2 + \big(1-\lambda_i(\omega,k)\big)|Y_i(\omega,k)|^2 \tag{6.4}$$

where, i is the subband number, k is the frame index, ω is the frequency bin index, $|\widehat{D}_i(\omega,k)|^2$ is the estimated subband noise power spectral density at frame k and frequency ω and $|Y_i(\omega,k)|^2$ is the noisy speech magnitude squared spectrum of subband noisy speech. Here, $\lambda_i(\omega,k)$ is a time and frequency dependent smoothing parameter, at subband i for frame, whose value depends on the noise changing rate.

In the recursive averaging technique, the smoothing parameter is chosen to be a sigmoid function which changes of the *a-posteriori* $\mathrm{SNR}_i(\omega,k)$, as

$$\lambda_i(\omega,k) = \frac{1}{1+\ e^{\left(-a(\mathrm{SNR}_i\,(\omega,k)-T)\right)}} \tag{6.5}$$

where the noise changing rate is affected by the parameter a in sigmoid function (6.5). The expression in (6.5) is obtained for the subband specific case for the sigmoid function as in (5.2). The value of a varies between the range 1 to 6 at constant value of T. The parameter T in (6.5) is the center-offset of the transition curve in sigmoid function and the value of T is around 4 to

5. The effect of a and T on the sigmoid function can be observed from Fig. 5.1, the simulation conditions of which is explained in Table 5.1 in Chapter 5.

The updation of subband noise estimate must be performed only in the absence of speech at the corresponding frequency bin. This can be accomplished by controlling the smoothing factor $\lambda_i(\omega, k)$ depending of the *a-posteriori* $\mathrm{SNR}_i(\omega, k)$ in i^{th} subband, as

$$\mathrm{SNR}_i(k) = 10\log_{10}\left(\frac{|Y_i(\omega,k)|^2}{\frac{1}{m}\sum_{p=1}^{m}|\widehat{D}_i(\omega,k-p)|^2}\right) \tag{6.6}$$

The denominator part of (6.6) is the average of the noise estimate of the previous m frames (numbering between 5 to 10) immediately before the frame k.

Theoretically the *a-posteriori* SNR should always be 1 when only noise is present and greater than 1 when both speech and noise are present. The progression of the noise estimation algorithm as given in (6.4) and (6.5) is given below.

i) If speech is present in frame k, the *a-posteriori* estimate $\mathrm{SNR}_i(\omega, k)$ will be large and therefore $\lambda_i(\omega, k) \approx 1$. Consequently, we will have $|\widehat{D}_i(\omega, k)|^2 \equiv |\widehat{D}_i(\omega, k-1)|^2$ according to (6.4). The noise update will cease and the noise estimate will remain the same as the previous frame's estimate.

ii) If speech is absent in frame k, the *a-posteriori* estimate $\mathrm{SNR}_i(\omega, k)$ will be small and therefore $\lambda_i(\omega, k) \cong 0$. As a result, $|\widehat{D}_i(\omega, k)|^2 \equiv |Y_i(\omega, k)|^2$ and the noise estimate will follow the power of the noisy spectrum in the absence of speech.

The main advantage of using the time-varying smoothing factor $\lambda_i(\omega, k)$, is that the noise estimation will adapt to different rates for the various frequency bins, depending on the estimate of the *a-posteriori* $\mathrm{SNR}_i(\omega, k)$ in the corresponding frequency-bins. Thus, the approach described above has the potential to effectively suppress the noise in each subband, separately, and is more suited for enhancement of speech degraded by non-stationary noise.

6.5. Improved Spectral Over-Subtraction Algorithm

The spectral over-subtraction algorithm, proposed by Berouti [20], is used for enhancement of speech degraded by stationary noise. As explained in Section 2.5.1, the spectral over-subtraction algorithm, proposed by Berouti [20] shows the superior result from the classical spectral subtraction method. In our algorithm, PMS-MBSS, the spectral over-subtraction algorithm is modified suitably and applied in each subband signal separately, which are obtained by decomposition of the input noisy speech using PM-SWPFB as described in section 6.3. The noise is estimated from each subband by using adaptive noise estimation approach and over-subtraction factor is adjusted for each subband signal. The improved spectral over-subtraction (I-SOS) algorithm as used in the proposed speech enhancement algorithm is explained below:

The noisy speech signal is decomposed into tempo-spectral stationary wavelet coefficients of multiple subbands by the PM-SWPFB. As, in (2.1), for the i^{th} noisy subband signal, i.e., the output of the i^{th} CB is given by

$$y_i(n) = h_i(n) * y(n) = s_i(n) + d_i(n) \tag{6.7}$$

where $s_i(n) = h_i(n) * s(n)$ is the output from the i^{th} CB filter when the input to the filterbank is clean speech only, and $d_i(n) = h_i(n) * d(n)$ is the corresponding output when the input is noise only.

Next, the improved spectral over-subtraction (I-SOS) algorithm is applied in each subband signal, separately. The noise is estimated from each subband signal by using adaptive noise estimation approach and over-subtraction factor is adjusted for each of the subband signals. Therefore, the estimate of the clean speech spectrum in the i^{th} subband signal is obtained by

$$|\hat{S}_i(\omega)|^2 = \begin{cases} |Y_i(\omega)|^2 - \alpha_i.|\widehat{D}_i(\omega)|^2, & \text{if } |\hat{S}_i(\omega)|^2 > \beta.|\widehat{D}_i(\omega)|^2 \\ \beta.|\widehat{D}_i(\omega)|^2 & \text{else} \end{cases} \tag{6.8}$$

Here, α_i is the subband specific over-subtraction factor, which is the function of the segmental SNR. The segmental SNR for each frame of the i^{th} subband signal can be calculated as

$$\text{SNR}_i\ (\text{dB}) = 10\log_{10}\left(\frac{\sum_{k=0}^{N_i-1}|Y_i(\omega)|^2}{\sum_{k=0}^{N_i-1}|\widehat{D}_i(\omega)|^2}\right) \tag{6.9}$$

where N_i is the number of samples in the each subband, k is the frame index and the value of $|\widehat{D}_i(\omega)|^2$ is estimated using (6.4). The subband specific over-subtraction can be calculated as

$$\alpha_i = \alpha_0 + (\mathrm{SNR}_i - \mathrm{SNR}_{\min})\left(\frac{\alpha_{\min}-\alpha_0}{\mathrm{SNR}_{\max}-\mathrm{SNR}_{\min}}\right), \text{ if } \mathrm{SNR}_{\min} \leq \mathrm{SNR}_i \leq \mathrm{SNR}_{\max} \qquad (6.10)$$

Here, for our proposed algorithm, PMS-MBSS, the typical values for (6.10) are, $\alpha_{\min} = 1$, $\alpha_{\max} = \alpha_0$, $\mathrm{SNR}_{\min} = 0$ dB, $\mathrm{SNR}_{\max} = 20$ dB and $\alpha_0 \approx 4$, .

6.6. Experimental Results and Performance Evaluation

This section presents the experimental results and performance evaluation of the proposed enhancement algorithm, PMS-MBSS, described in this chapter, as well as the comparison with other spectral subtractive-type algorithms. The noisy speech samples have been taken from NOIZEUS speech corpus [76] consisting of 30 phonetically balanced sentences belonging to six speakers, three male and three female. The sampling frequency for this corpus is 8 kHz, and samples are quantized linearly using 16 bits. The NOIZEUS corpus comes with non-stationary noises at different levels of SNR. A total of four different utterances, from NOIZEUS corpus, are used in our evaluation. Every sentence has a silence segment in the beginning that lasts more than 0.25 ms.

Eight different types of noises including one stationary white Gaussian noise have been used for simulation and evaluation purpose. The seven real-world non-stationary noises taken are car, train, restaurant, babble, airport, street, and exhibition noise. The performance of the proposed speech enhancement system is tested on these noisy speech samples.

In our enhancement experiments, we normalize samples of each of the sentence files to be between -1.0 and +1.0. The frame size is chosen to be 256 samples, which contains a time window of 32ms, with 50% overlapping. The sinusoidal Hamming window with size 256 samples is applied to the noisy signal. The noise estimate is updated adaptively and continuously, using the smoothing parameter (6.5). For calculation of smoothing parameter, the value of a and T is chosen to be 4 and 5 (as Fig. 5.1), respectively in sigmoid function (6.5). The proposed algorithm, PMS-MBSS, is compared with the basic spectral subtraction (BSS), spectral over-subtraction (SOS), and multi-band spectral subtraction (MBSS) algorithm.

6.6.1 Selection of Wavelet Filter

The decomposition of stationary wavelet packet filterbank (SWPFB) has been explained in detail in Section 3.7 of Chapter 3. For the wavelet based signal processing, the choice of mother wavelet function and the corresponding wavelet filter is important for the frequency selectivity. In addition, the computational complexity of the stationary wavelet packet filterbank (SWPFB) is directly dependent on the length of the wavelet filter. To put differently, the use of a longer wavelet filter will provide a better frequency resolution at the cost of heavy computational complexity.

The wavelet coefficients, thus obtained, are processed spectrally in each subband. We may encounter some aliased noises in side-lobes because of the insufficient length of the wavelet filter. Therefore, a sufficient stop-band attenuation of the wavelet filter is required and the longer wavelet filters are needed. In our proposed algorithm, the orthogonal wavelet filters including Daubechies ('DbN'), Symlets ('symN'), and Cioflets ('coifN') are considered in the PM-SWPFB, and their relative performance compared [16].

In order to select an appropriate wavelet filter for the proposed speech enhancement algorithm from these wavelet filters, an experiment is performed, the result of which is listed in Table 6.2.

TABLE 6.2.
THE EXPERIMENTAL RESULTS FOR SELECTION OF WAVELET FILTER.

Wavelet filter type	**Daubechies (Db N)**					**Symlet (symN)**			**Coiflets (coif N)**	
	Db 4	**Db8**	**Db10**	**Db12**	**Db14**	**S12**	**S14**	**S16**	**C3**	**C5**
Filter length	8	16	20	24	28	24	28	32	18	30
ER	5.22	5.23	5.29	5.28	5.27	5.29	5.25	5.29	5.26	5.27
CPU time (Sec.)	4.48	4.66	4.79	4.72	4.95	4.77	5.03	5.24	5.17	5.18
ER/CPU time	1.16	1.12	1.10	1.12	1.06	1.11	1.04	1.00	1.02	1.01

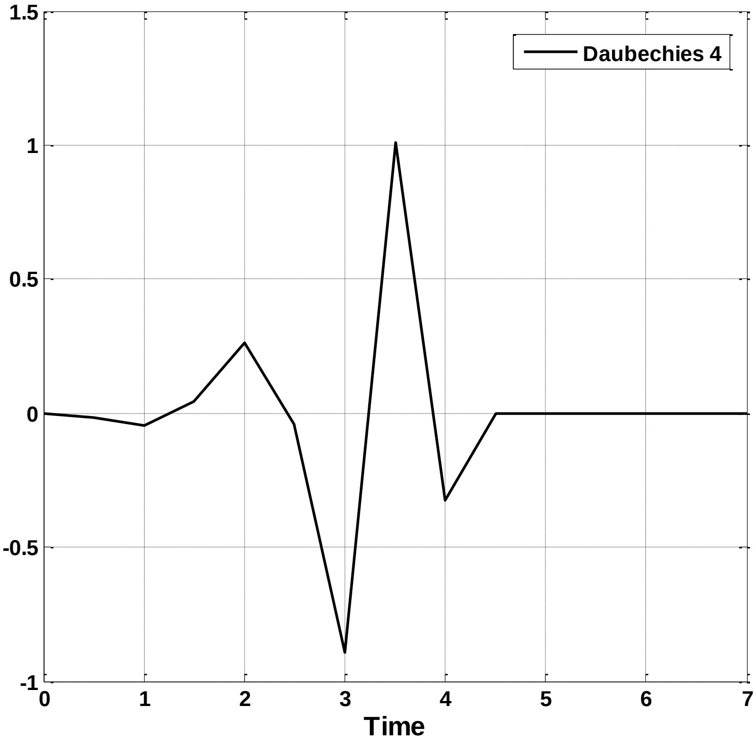
The enhancement rate (ER), as tabulated in Table 6.2 is defined as,

$$\mathrm{ER} = \frac{\sum_{i=1}^{M}\{\mathrm{SNR}[\hat{s}_i(n)] - \mathrm{SNR}[x_i(n)]\}}{M} \tag{6.11}$$

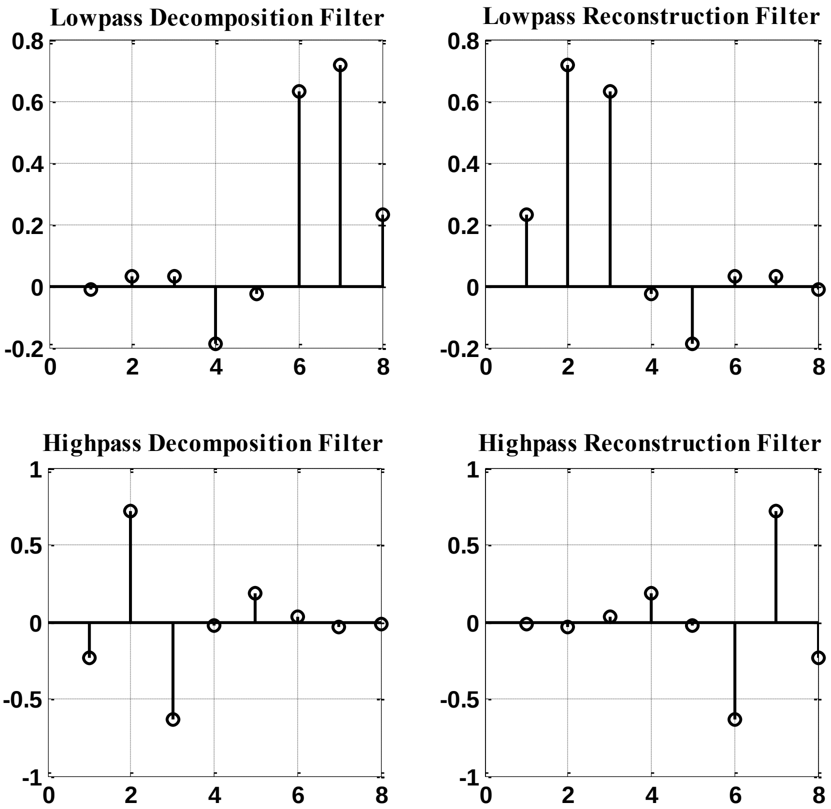
where M denotes the number of test sentences used in the experiment. Here, $\mathrm{SNR}[\hat{s}_i(n)]$, and $\mathrm{SNR}[x_i(n)]$ denote the SNR of the i^{th} enhanced speech signal and the original speech signal, respectively. In our experiment, 4 $(M = 4)$ speech sentences degraded by different real-world noises and white Gaussian noise are used.

Considering the enhancement rate as well as the computational complexity given in Table 6.2, the 4-point Daubechies (Db) wavelet filter with length 8, has the best ER/CPU time ratio, and has been taken for the implementation of the proposed speech enhancement algorithm. This preserves sufficient frequency selectivity as well as maintains the time domain resolution for each critical band wide subband signal.

The filterbanks are implemented using the high pass filter (HPF) and low pass filter (LPF) with the orthonormal wavelet, Daubechies (DbN) family wavelet [25]. The wavelet function, and the filter coefficients for Db 4 wavelet is shown in Fig. 6.5 (a), and (b), respectively.



(a)



(b)

Fig. 6.5: (a) Daubechies wavelet (Db 4), and (b) impulse response of the corresponding low pass and high pass filter.

The PM-SWPFB is first applied to decompose the noisy speech signal into 17 subband signals (Fig. 6.3). The filterbanks are implemented using the high pass filter (HPF) and low pass filter (LPF) with the Daubechies (Db4) family wavelet, which divides the whole band into the perceptually motivated subbands. In the first level decomposition, scaling space and wavelet space will be decomposed into two subbands, which correspond to the frequency ranges of 0− 2 kHz and 2− 4 kHz. This operation is repeated to obtain the filterbank structure with highest level of decomposition of five.

Fig. 6.6 shows the temporal waveforms of noisy speech signal decomposed by PM-SWPFB into 1st to 8th CBs and the enhanced speech of these subbands for speech sentence sp10 pronounced by male speaker and degraded by car noise of SNR 10 dB, from NOIZEUS [76].

From the left part of this figure it can be seen that the noise distributes quite differently across these subbands. The enhanced speech components in these subbands are shown in right half of this figure.

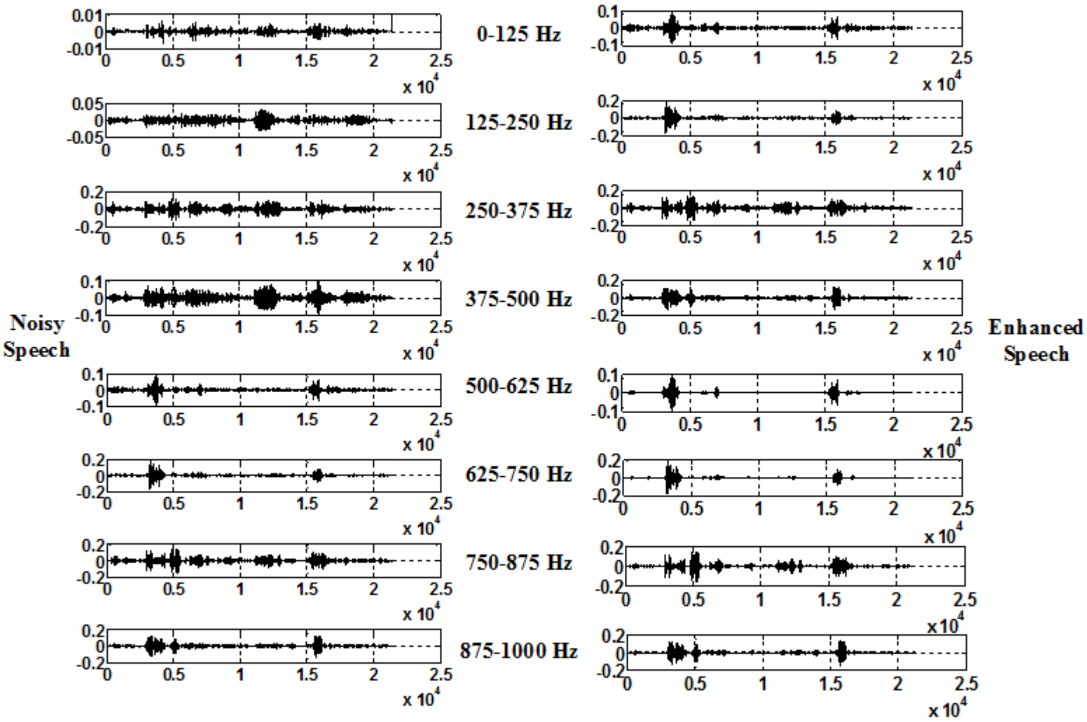


Fig. 6.6: Temporal waveform of subbands (obtained after decomposing noisy speech, which is degraded by car noise at 10 dB SNR, by PM-SWPFB) and enhanced subbands speech by I-SOS.

6.6.2 Objective Measure

The input SNR vs. output SNR of the proposed speech enhancement algorithm for seven types of real-world noises and a computer generated white Gaussian noises have been shown in Table 6.3. The amount of noise reduction in various background noise level conditions is usually measured with the SNR improvement which is given by the difference between input and output SNR. The SNR improvement of, PMS-MBSS, over BSS, SOS and MBSS algorithm is shown in Table 6.4. In the non-stationary noise environment, SNR of enhanced speech is not a sufficient objective measure of speech quality. In addition, three other objective quality measures, i.e. SegSNR, ISD, and PESQ, are used to evaluate the performance of proposed algorithm. Table 6.3, further, shows the output SegSNR and ISD values for various real-world noises and white Gaussian noise at different SNR levels. In Table 6.4, we shows the SegSNR improvement in comparison over BSS, SOS and MBSS. Next, Fig. 6.7 shows the performance of PMS-MBSS for various real-world noises and white Gaussian noise at different SNR levels and Fig. 6.8 shows the performance improvement of PMS-MBSS at various real-world noises and white Gaussian noise at different SNR levels over SOS and MBSS algorithms.

PESQ is an objective quality measure designed to predict the subjective opinion score of a degraded audio sample and it is recommended by ITU-T for speech quality assessment [75]. The larger value of PESQ indicates a better subjective quality. Table 6.5 shows the PESQ improvement of PMS-MBSS over BSS, SOS and MBSS. It can be observed from the table that the PESQ score of PMS-MBSS shows the best results in comparison to MBSS, BSS and is comparable with SOS.

In order to analyze the time-frequency distribution of the enhanced speech, speech spectrogram constitutes a well-suited tool to give accurate information about remnant noise and speech distortion. It consists of Fourier transforms of overlapping, and windowed frames and displays the distribution of energy in time and frequency. Fig. 6.10 shows the spectrogram obtained with the proposed method for the speech sentence (sp10) degraded by car noise, train noise, babble noise, restaurant noise, airport noise, street noise, exhibition noise, and white noise, respectively at 10 dB SNR.

Fig. 6.11 shows the temporal waveforms obtained with the proposed algorithm for the speech sentence (sp10) degraded by car noise, train noise, babble noise, restaurant noise, airport noise, street noise, exhibition noise, and white noise, respectively at 10 dB SNR. Fig. 6.12 - Fig. 6.15 shows the temporal waveforms and spectrogram obtained with the proposed algorithm for the male speech sentences (sp10, sp6, and sp1) and female speech sentences (sp12) degraded by car noise at 10 dB SNR in comparison to BSS, SOS, and MBSS algorithms.

It can be seen from Fig.6.10 - Fig.6.15 that the speech enhanced by the proposed algorithm is better compared to MBSS, BSS and is also comparable to SOS. The musical structure of the remnant noise is also found to be suppressed more in comparison to MBSS, whereas the reduction is comparable in respect to SOS.

For the case with sp1 sentence pronounced by male speaker our results show the performance even better than SOS, as noticed from Fig. 6.14.

6.6.3 Subjective Listening Tests

Fig. 6.8 shows a scatter plot of MOS score vs. PESQ scores, for PMS-MBSS, for various types of real-world noises and the computer generated white Gaussian noise. A straight line with the slope of one is provided as a reference. It is clear from the figure that the PESQ improvement score and MOS score at 5 dB SNR and above 5 dB SNR is well correlated in all conditions. Table 6.5 shows the MOS score results of PMS-MBSS over SOS and MBSS. It can be observed from the table that the MOS score results of PMS-MBSS are comparable with SOS and is gives better result in comparison to MBSS.

Experimental results using subjective and objective quality measurement test results have shown the superiority of the proposed speech enhancement algorithm to the other popular speech enhancement algorithms, such as, MBSS, BSS and SOS. The following conclusions can be extracted about experimental results and objective quality measurement test results.

i) The stationary wavelet packet decomposition (SWPD) provides sufficient numbers of sample points, i.e. sufficient frequency resolutions, for designing a PM-SWPFB that closely matches with critical band structure of human auditory system.

ii) The PM-SWPFB provides improved speech quality and low computational complexity.

iii) I-SOS algorithm effectively reduces remnant noise and background noise, when of α_i is set as per (6.10) and the value of $\beta = 0.03$. Parameters (smoothing factor) should be selected carefully, as there is a trade-off between noise removal and signal distortion in I-SOS algorithm.

iv) At high SNRs (5, 10, and 15 dB) the proposed speech enhancement algorithm, PMS-MBSS, provides the best results. Although, the SNR value of SOS algorithm is higher than that of proposed algorithm, the SegSNR and ISD results of the PMS-MBSS is better, than BSS, MBSS and SOS, as given in Table 6.3 and Table 6.4. It can also be seen from the PESQ score results and MOS score results that the PMS-MBSS gives better result than BSS, MBSS and SOS. This has been presented in Table 6.5 and Fig. 6.9.

v) After carefully analyzing the signal waveforms and spectrogram (Fig. 6.10 - Fig. 6.15), it can be observed that the proposed algorithm reduces the musical structure of the remnant noise more than BSS, MBSS, whereas the performance is comparable to SOS. For sp1 sentence pronounced by male our results are the best in comparison to even SOS. However, the proposed speech enhancement algorithm provides a good noise removal and less signal distortion is obtained even at low SNRs. Therefore, speech enhanced with the proposed algorithm is more pleasant with speech distortion remaining below the acceptable level.

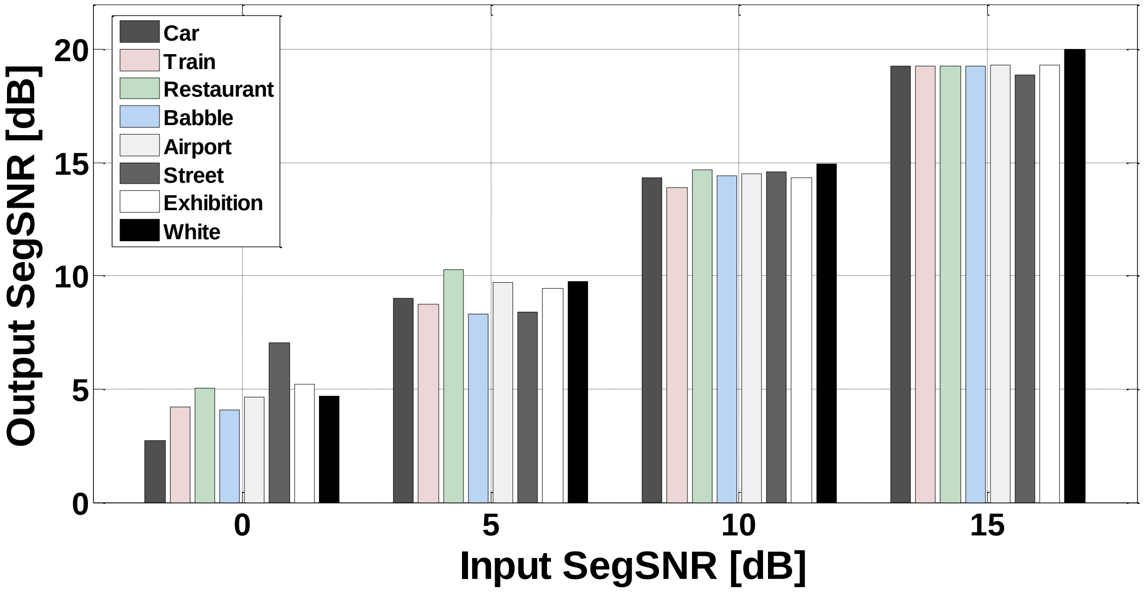


Fig. 6.7: Output SegSNR of PMS-MBSS for car, train, restaurant, babble, airport, street, exhibition, and white noises.

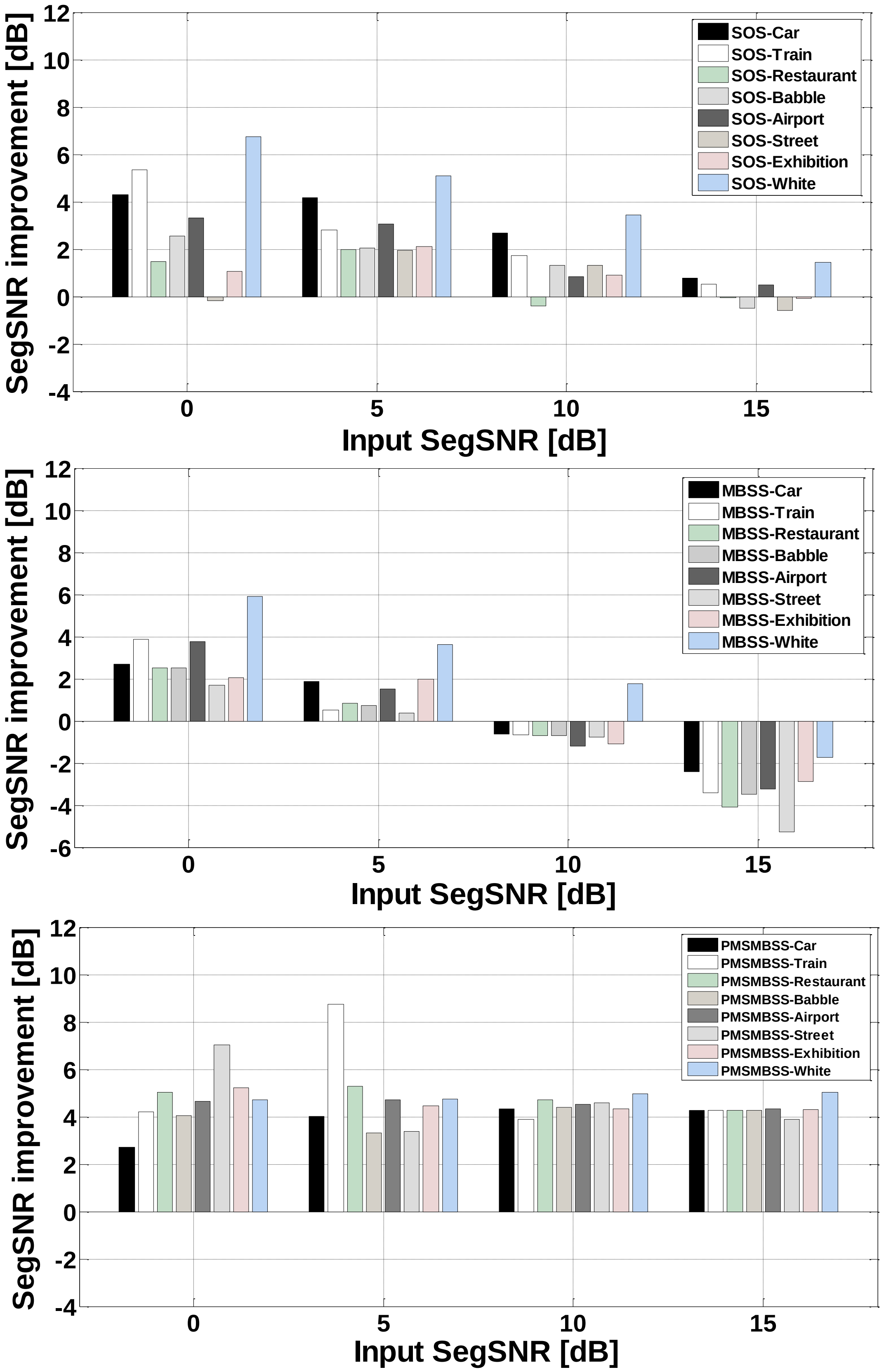


Fig. 6.8: SegSNR improvement of PMS-MBSS over SOS and MBSS for car, train, restaurant, babble, airport, street, exhibition, and white noises.

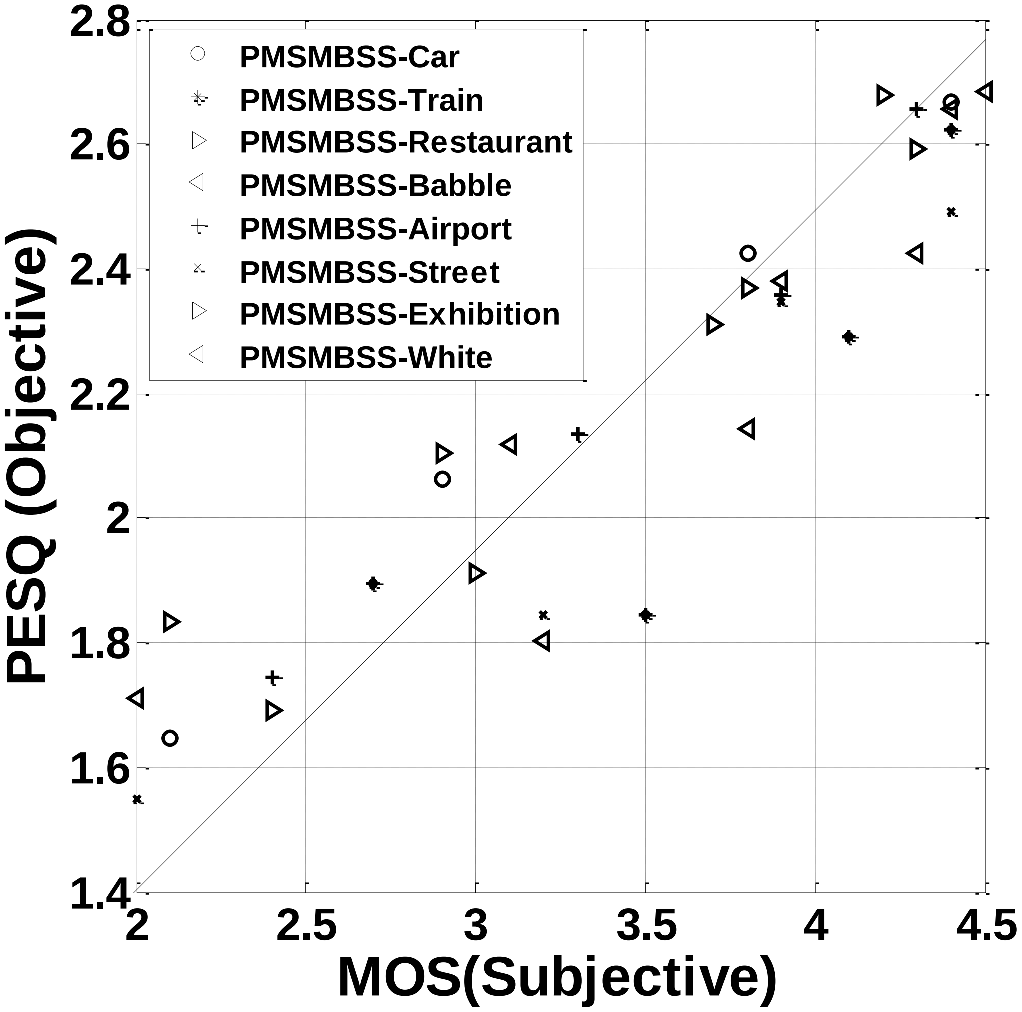


Fig. 6.9: Scattered plot of PESQ score vs. mean MOS score of PMS-MBSS for car, train, restaurant, babble, airport, street, exhibition, and white noises.

.

TABLE 6.3.

OBJECTIVE MEASURES OBTAINED WITH THE PROPOSED ALGORITHM FOR THE VARIOUS NOISE TYPES IN TERMS OF OUTPUT SNR, OUTPUT SEGSNR, AND ISD.

Noise Type	**Enhancement Algorithms**	**SNR (dB)**				**SegSNR (dB)**				**ISD**			
		0 dB	**5 dB**	**10 dB**	**15 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**
Car	**BSS**	0.95	1.40	1.59	1.71	0.86	1.17	1.28	1.33	0.84	0.29	0.08	0.212
	SOS	4.39	9.42	12.89	16.01	4.30	9.19	12.68	15.78	2.29	1.48	0.90	0.78
	MBSS	2.75	7.05	9.61	12.85	2.68	6.86	9.39	12.59	2.27	1.43	0.89	0.76
	PMS-MBSS	**2.74**	**5.75**	**9.86**	**14.53**	**2.71**	**9.02**	**14.34**	**19.28**	**2.43**	**1.44**	**0.85**	**0.78**
Train	**BSS**	1.24	1.43	1.59	1.67	1.12	1.20	1.29	1.32	0.17	0.25	0.30	0.099
	SOS	5.54	8.04	11.91	15.74	5.34	7.81	11.73	15.53	2.15	1.41	1.13	0.89
	MBSS	4.07	5.73	9.55	11.78	3.88	5.53	9.35	11.59	2.15	1.40	1.34	0.86
	PMS-MBSS	**2.72**	**5.70**	**9.81**	**14.53**	**4.22**	**8.76**	**13.89**	**19.26**	**2.01**	**1.63**	**1.06**	**0.86**
Restaurant	**BSS**	2.07	1.59	1.84	1.64	1.75	1.27	1.37	1.30	0.48	.002	0.125	0.066
	SOS	1.72	7.20	9.81	15.13	1.48	7.00	9.59	14.95	1.42	1.02	0.82	2.79
	MBSS	2.66	6.02	9.54	11.18	2.52	5.84	9.29	10.90	1.11	1.27	0.65	2.70
	PMS-MBSS	**2.69**	**5.74**	**9.81**	**14.50**	**5.02**	**10.3**	**14.70**	**19.26**	**1.31**	**0.72**	**29.14**	**0.36**
Babble	**BSS**	1.47	1.52	1.63	1.74	1.19	1.21	1.29	1.35	0.35	0.42	0.073	0.023
	SOS	2.74	7.34	11.52	14.74	2.57	7.05	11.33	14.51	1.61	1.49	2.12	0.72
	MBSS	2.66	5.95	9.54	11.81	2.52	5.74	9.31	11.52	1.22	1.21	1.42	0.43
	PMS-MBSS	**2.66**	**5.75**	**9.90**	**14.53**	**4.06**	**8.32**	**14.41**	**19.26**	**1.41**	**1.33**	**0.73**	**0.56**
Airport	**BSS**	1.66	1.52	1.62	1.72	1.43	1.23	1.30	1.33	0.59	0.06	0.016	0.14
	SOS	3.60	8.30	11.04	15.68	3.33	8.06	10.85	15.48	1.80	0.91	1.06	0.75
	MBSS	3.93	6.75	9.06	12.04	3.78	6.53	8.81	11.77	1.28	0.90	1.05	0.77
	PMS-MBSS	**4.81**	**9.88**	**14.66**	**19.48**	**4.66**	**9.71**	**14.52**	**19.32**	**14.8**	**1.01**	**0.92**	**0.41**
Street	**BSS**	2.49	1.51	1.66	1.55	1.71	1.25	1.31	1.29	0.13	0.51	0.067	0.14
	SOS	0.02	7.14	11.17	14.61	-0.17	6.96	11.33	14.41	1.04	1.46	5.15	1.10
	MBSS	1.88	5.60	9.42	9.93	1.71	5.39	9.22	9.73	1.04	1.11	5.14	0.97
	PMS-MBSS	**7.00**	**8.50**	**14.55**	**19.11**	**7.03**	**8.39**	**14.60**	**18.89**	**0.98**	**1.36**	**0.78**	**1.09**
Exhibition	**BSS**	2.08	1.63	1.66	1.74	2.01	1.34	1.35	1.37	0.11	0.10	0.38	0.07
	SOS	1.29	7.31	11.13	15.12	1.07	7.10	10.9	14.91	1.69	1.10	1.32	0.60
	MBSS	2.24	7.18	9.09	12.36	2.05	6.99	8.92	12.13	1.68	1.06	1.21	0.51
	PMS-MBSS	**5.33**	**9.67**	**14.47**	**19.50**	**5.22**	**9.47**	**14.33**	**19.30**	**1.26**	**0.95**	**1.31**	**0.71**
White	**BSS**	1.42	1.59	1.70	1.78	1.18	1.31	1.34	1.38	0.60	0.38	0.25	0.214
	SOS	6.98	10.3	13.65	16.67	6.75	10.1	13.43	16.45	2.46	1.94	1.46	0.99
	MBSS	6.10	8.80	11.93	13.46	5.90	8.63	11.77	13.26	2.40	1.88	1.45	0.99
	PMS-MBSS	**4.84**	**9.97**	**15.18**	**20.20**	**4.70**	**9.74**	**14.97**	**20.03**	**2.38**	**1.93**	**1.45**	**0.96**

TABLE 6.4.

OBJECTIVE MEASURES OBTAINED WITH THE PROPOSED ALGORITHM FOR THE VARIOUS NOISE TYPES IN TERMS OF SNR IMPROVEMENT, AND SEGSNR IMPROVEMENT.

Noise Type	**Enhancement Algorithms**	**SNR Improvement (dB)**				**SegSNR Improvement (dB)**			
		0 dB	**5 dB**	**10 dB**	**15 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**
Car	**BSS**	0.95	-3.6	-8.41	-13.29	0.86	-3.83	-8.72	-13.67
	SOS	4.39	4.42	2.89	1.01	4.30	4.19	2.68	0.78
	MBSS	2.75	2.05	-0.39	-2.15	2.68	1.86	-0.61	-2.41
	PMS-MBSS	**2.74**	**0.75**	**-0.14**	**-0.47**	**2.71**	**4.02**	**4.34**	**4.28**
Train	**BSS**	1.24	-3.57	-8.41	-13.33	1.12	-3.8	-8.71	-13.68
	SOS	5.54	3.04	1.91	.74	5.34	2.81	1.73	0.53
	MBSS	4.07	0.73	-0.45	-3.22	3.88	0.53	-0.65	-3.41
	PMS-MBSS	**2.72**	**0.70**	**-0.19**	**-0.47**	**4.22**	**8.76**	**3.89**	**4.26**
Restaurant	**BSS**	2.07	-3.41	-8.16	-13.36	1.75	-3.73	-8.63	-13.7
	SOS	1.72	2.20	0.19	0.13	1.48	2.00	-0.41	-0.05
	MBSS	2.66	1.02	-0.46	-3.82	2.52	0.84	-0.71	-4.1
	PMS-MBSS	**2.69**	**0.74**	**-0.19**	**-0.50**	**5.02**	**5.3**	**4.70**	**4.26**
Babble	**BSS**	1.47	-3.48	-8.37	-13.26	1.19	-3.79	-8.71	-13.65
	SOS	2.74	2.34	1.52	-0.26	2.57	2.05	1.33	-0.49
	MBSS	2.66	0.95	-0.46	-3.19	2.52	0.74	-0.69	-3.48
	PMS-MBSS	**2.66**	**0.75**	**-0.1**	**-0.47**	**4.06**	**3.32**	**4.41**	**4.26**
Airport	**BSS**	1.66	-3.48	-8.38	-13.28	1.43	-3.77	-8.7	-13.67
	SOS	3.60	3.30	1.04	0.68	3.33	3.06	0.85	0.48
	MBSS	3.93	1.75	-0.94	-2.96	3.78	1.53	-1.19	-3.23
	PMS-MBSS	**4.81**	**4.88**	**4.66**	**4.48**	**4.66**	**4.71**	**4.52**	**4.32**
Street	**BSS**	2.49	-3.49	-8.34	-13.45	1.71	-3.77	-8.69	-13.71
	SOS	0.02	2.14	1.17	-0.39	-0.17	1.96	1.33	-0.59
	MBSS	1.88	0.60	0.58	-5.07	1.71	0.39	-0.78	-5.27
	PMS-MBSS	**7.00**	**3.50**	**4.55**	**4.11**	**7.03**	**3.39**	**4.60**	**3.89**
Exhibition	**BSS**	2.08	-3.37	-8.38	-13.26	2.01	-3.66	-8.65	-13.63
	SOS	1.29	2.31	1.13	0.12	1.07	2.10	0.9	-0.09
	MBSS	2.24	2.18	-0.91	-2.64	2.05	1.99	-1.08	-2.87
	PMS-MBSS	**5.33**	**4.67**	**4.47**	**4.50**	**5.22**	**4.47**	**4.33**	**4.30**
White	**BSS**	1.42	-3.41	-8.3	-13.22	1.18	-3.69	-8.66	-13.62
	SOS	6.98	5.30	2.63	1.67	6.75	5.1	3.43	1.45
	MBSS	6.10	3.80	1.93	-1.54	5.90	3.63	1.77	-1.74
	PMS-MBSS	**4.84**	**4.97**	**5.18**	**5.20**	**4.70**	**4.74**	**4.97**	**5.03**

TABLE 6.5.

RESULTS OF NOISE REDUCTION SPEECH QUALITY TEST.

Noise Type	Enhancement Algorithms	PESQ Score				MOS Score			
		0 dB	**5 dB**	**10 dB**	**15 dB**	**0 dB**	**5 dB**	**10 dB**	**15 dB**
Car	**BSS**	1.749	1.925	2.154	2.213				
	SOS	1.622	2.163	2.579	2.831	2.5	3.3	4	4.4
	MBSS	1.496	1.982	2.259	2.602	2.2	3	3.8	4.3
	PMS-MBSS	**1.648**	**2.062**	**2.426**	**2.667**	**2.1**	**2.9**	**3.8**	**4.4**
Train	**BSS**	1.873	1.666	2.079	2.156				
	SOS	1.720	1.888	2.394	2.730	2.3	3.2	4	4.5
	MBSS	1.513	1.696	2.129	2.382	2.2	3	3.9	4.3
	PMS-MBSS	**1.895**	**1.845**	**2.291**	**2.623**	**2.7**	**3.5**	**4.1**	**4.4**
Restaurant	**BSS**	1.682	1.843	2.002	2.165				
	SOS	1.785	2.157	2.362	2.811	2.6	3.5	4	4.4
	MBSS	1.842	2.062	2.367	2.603	2.4	3.2	4	4.3
	PMS-MBSS	**1.692**	**2.105**	**2.310**	**2.679**	**2.4**	**2.9**	**3.7**	**4.2**
Babble	**BSS**	1.481	1.924	2.110	2.215				
	SOS	1.903	2.209	2.562	2.699	2	3	3.9	4.4
	MBSS	1.812	2.208	2.394	2.650	1.8	2.8	3.9	4.4
	PMS-MBSS	**1.711**	**2.119**	**2.380**	**2.656**	**2**	**3.1**	**3.9**	**4.4**
Airport	**BSS**	1.407	1.939	2.092	2.204				
	SOS	1.891	2.222	2.476	2.836	2.6	3.4	4	4.2
	MBSS	1.790	2.106	2.323	2.681	1.7	2.9	3.6	4.4
	PMS-MBSS	**1.745**	**2.134**	**2.357**	**2.655**	**2.4**	**3.3**	**3.9**	**4.3**
Street	**BSS**	1.511	1.833	2.045	2.018				
	SOS	1.578	2.013	2.406	2.536	2.4	2.9	3.8	4.4
	MBSS	1.592	1.933	2.249	2.213	1.9	2.8	3.5	4.3
	PMS-MBSS	**1.548**	**1.845**	**2.346**	**2.492**	**2**	**3.2**	**3.9**	**4.4**
Exhibition	**BSS**	1.721	1.655	2.109	2.127				
	SOS	1.799	2.004	2.267	2.694	2.4	3.2	3.8	4.2
	MBSS	1.527	1.977	1.968	2.517	2	2.9	3.8	4.4
	PMS-MBSS	**1.835**	**1.913**	**2.370**	**2.593**	**2.1**	**3**	**3.8**	**4.3**
White	**BSS**	1.663	1.957	2.087	2.151				
	SOS	1.912	2.232	2.483	2.800	3.0	3.7	4.2	4.4
	MBSS	1.655	1.971	2.303	2.563	3	3.7	4.3	4.5
	PMS-MBSS	**1.803**	**2.144**	**2.426**	**2.683**	**3.2**	**3.8**	**4.3**	**4.5**

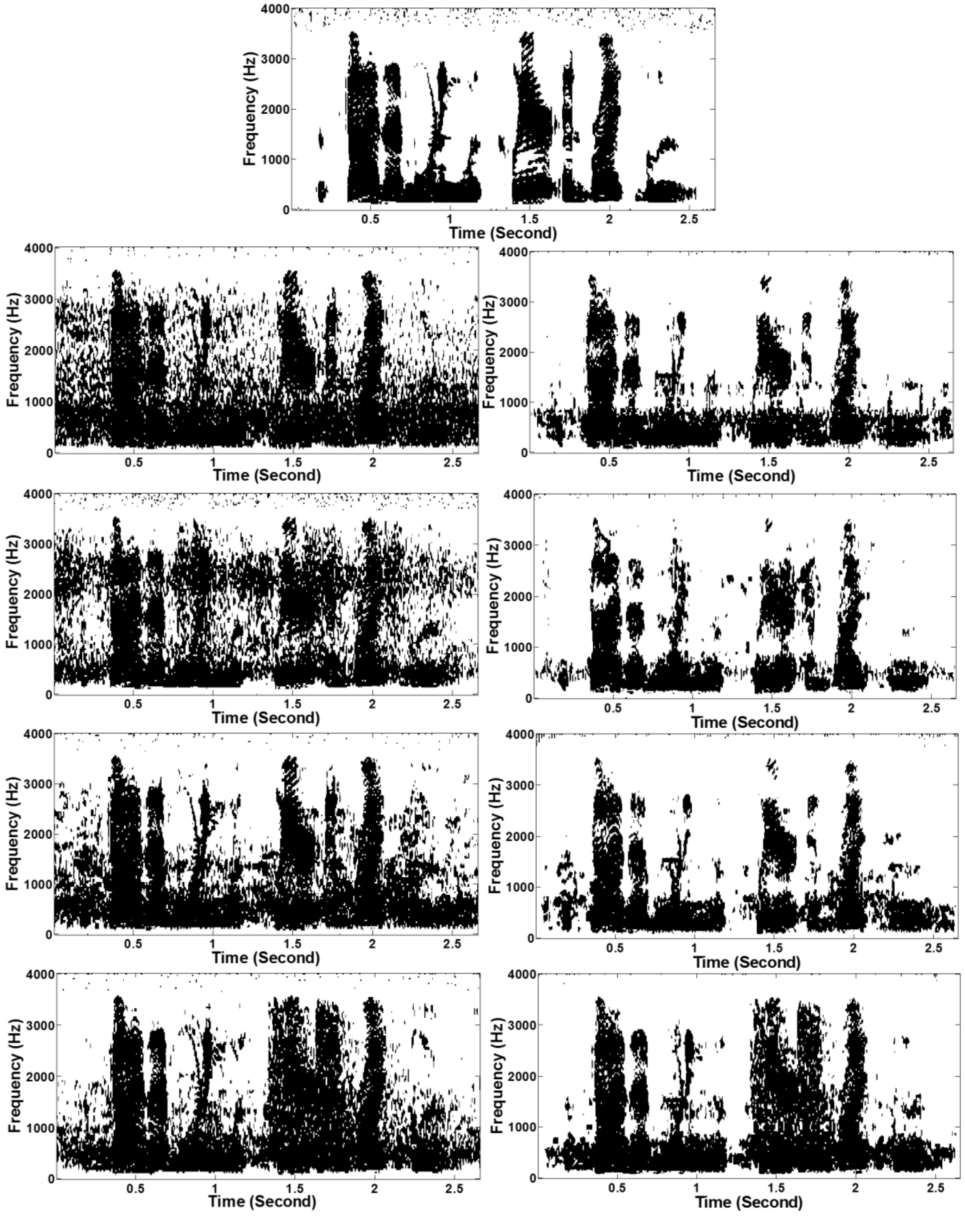


Fig. 6.10 (I): Speech spectrogram of *sp10.wav* utterance, "The sky that morning was clear and bright blue," by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech; (b, d, f, h) speech degraded by car noise, train noise, babble noise, and restaurant noise, respectively (10 dB SNR); and (c, e, g, i) corresponding enhanced speech.

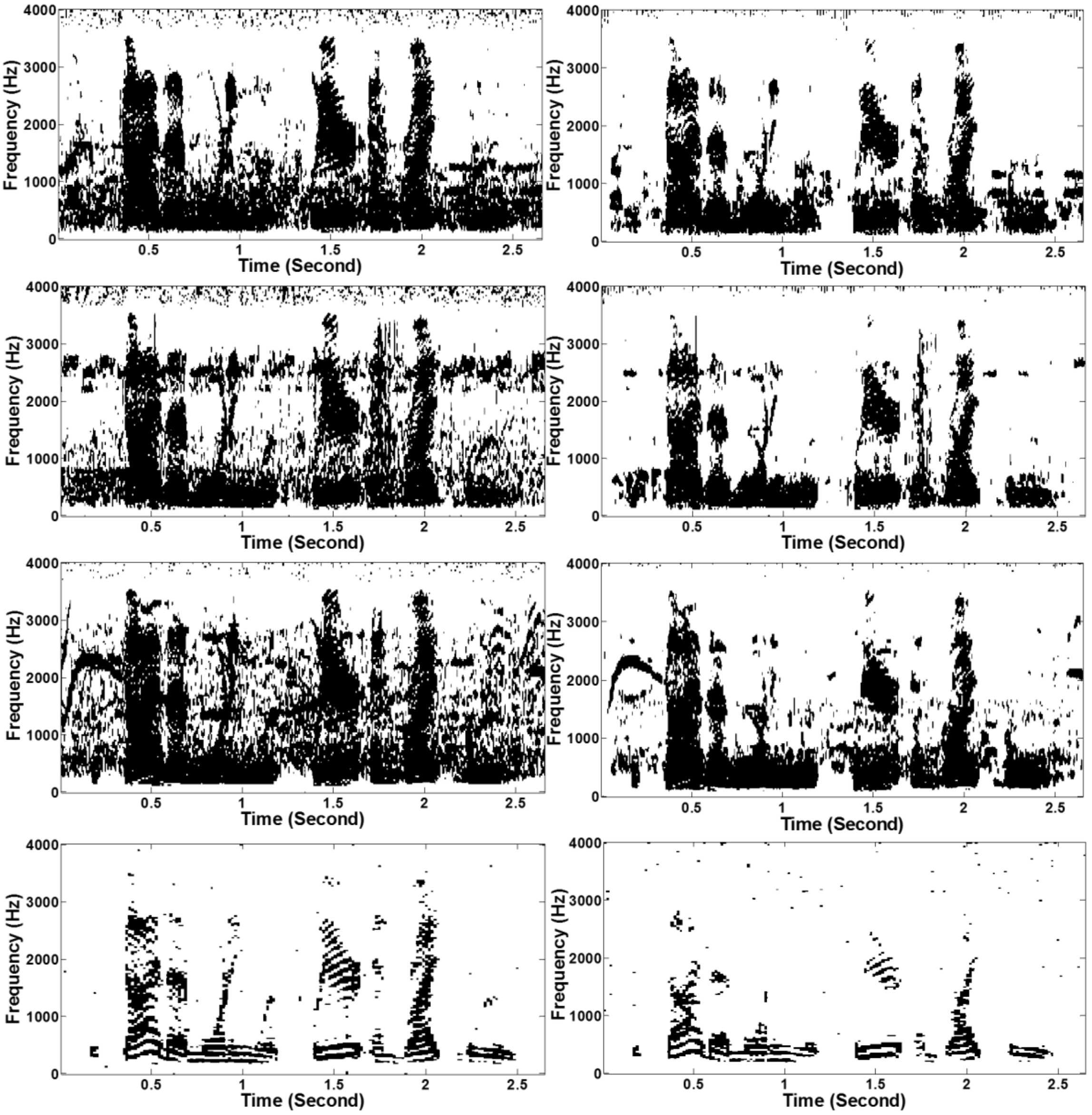


Fig. 6.10 (II): Speech spectrogram of *sp10.wav* utterance, "The sky that morning was clear and bright blue," by a male speaker from the NOIZEUS corpus (From top to bottom): (j, l, n, p) speech degraded by airport noise, street noise, exhibition noise, and white noise respectively (10 dB SNR); and (k, m, o, q) corresponding enhanced speech.

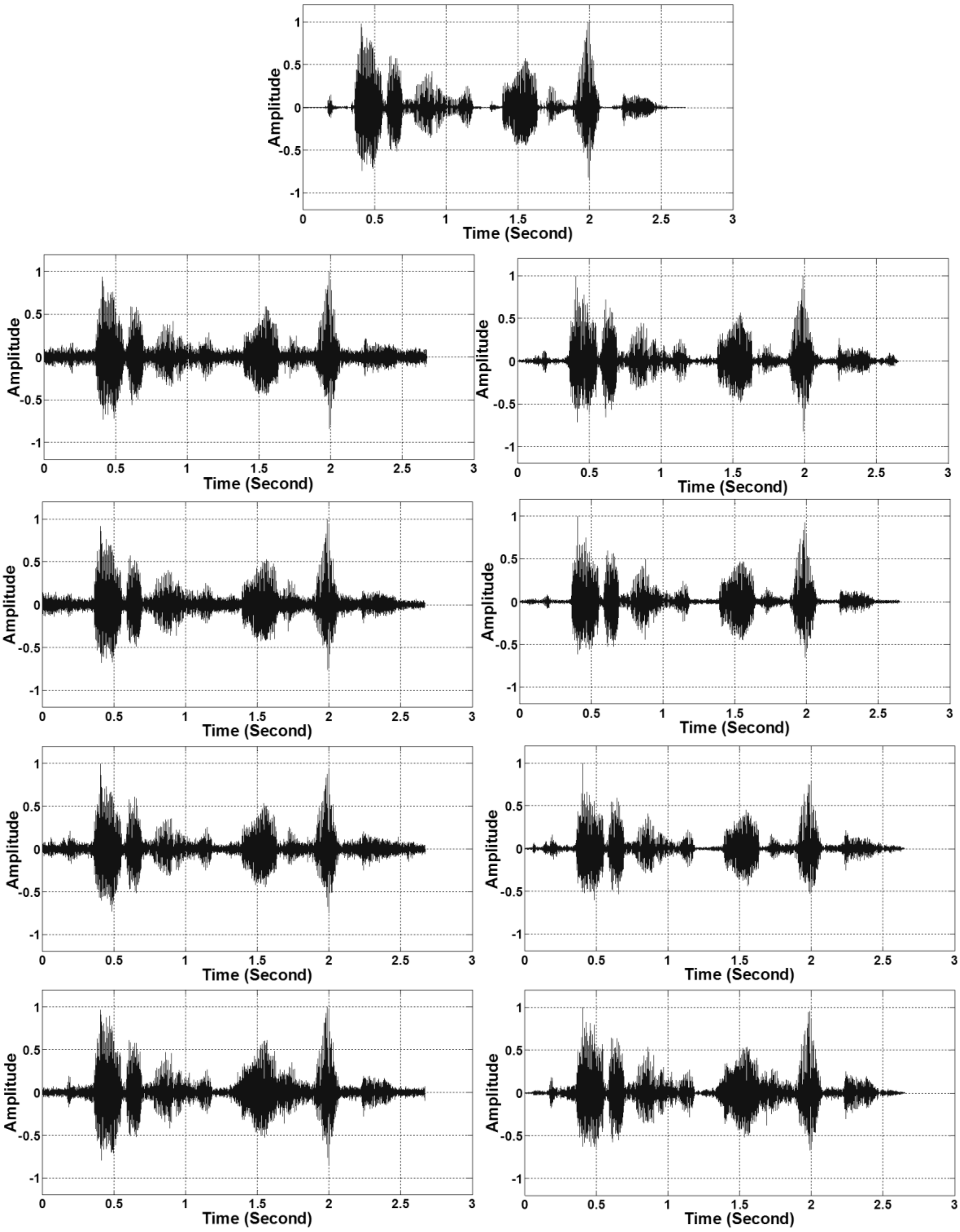


Fig. 6.11(I): Temporal waveforms of *sp10.wav* utterance, "The sky that morning was clear and bright blue," by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech; (b, d, f, h) speech degraded by car noise, train noise, babble noise and restaurant noise, respectively (10 dB SNR); and (c, e, g, i) corresponding enhanced speech.

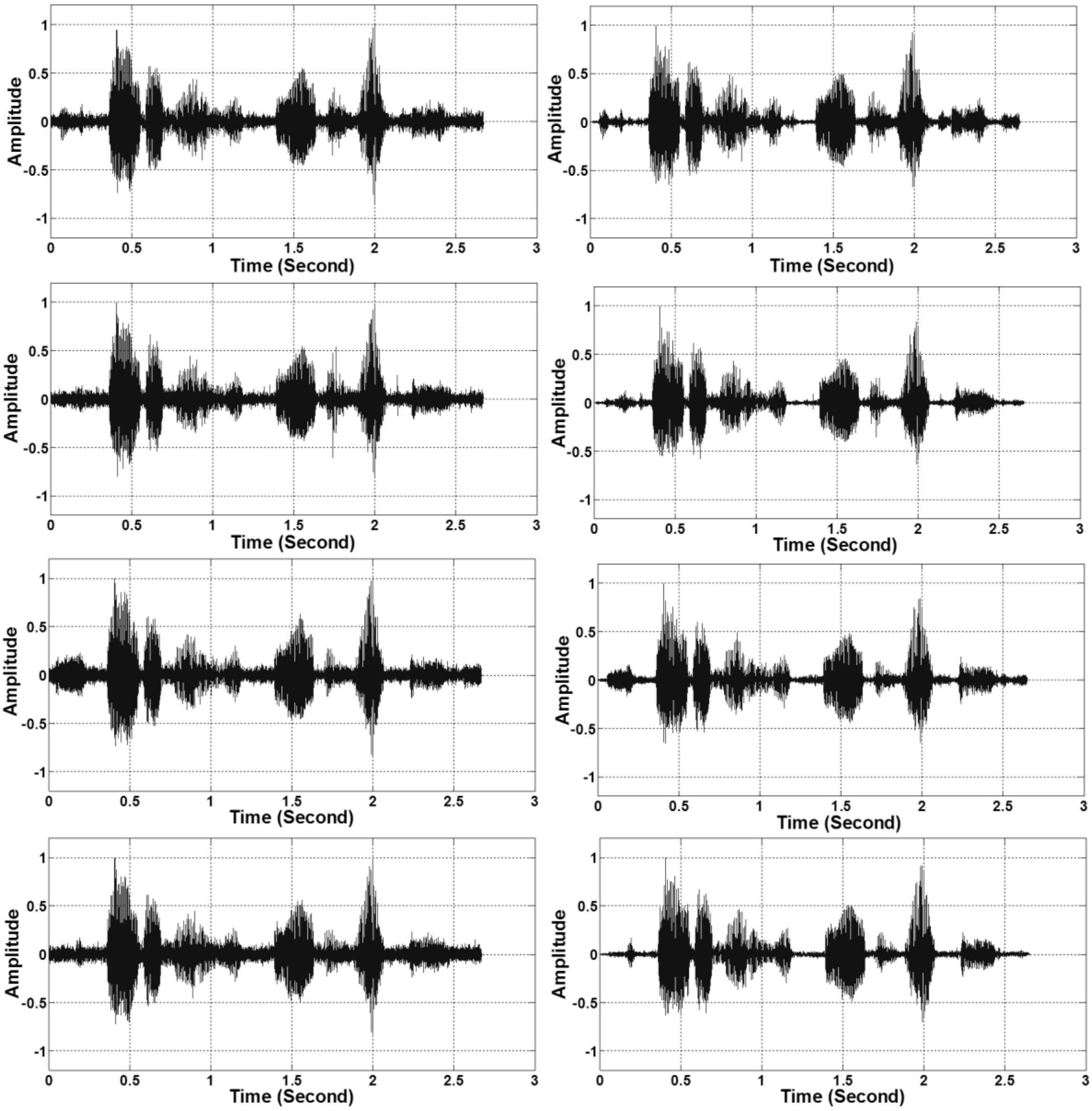


Fig. 6.11(II): Temporal waveforms of *sp10.wav* utterance, "The sky that morning was clear and bright blue," by a male speaker from the NOIZEUS corpus (From top to bottom (j, l, n, p) speech degraded by airport noise, street noise, exhibition noise, and white noise, respectively (10 dB SNR); and (k, m, o, q) corresponding enhanced speech.

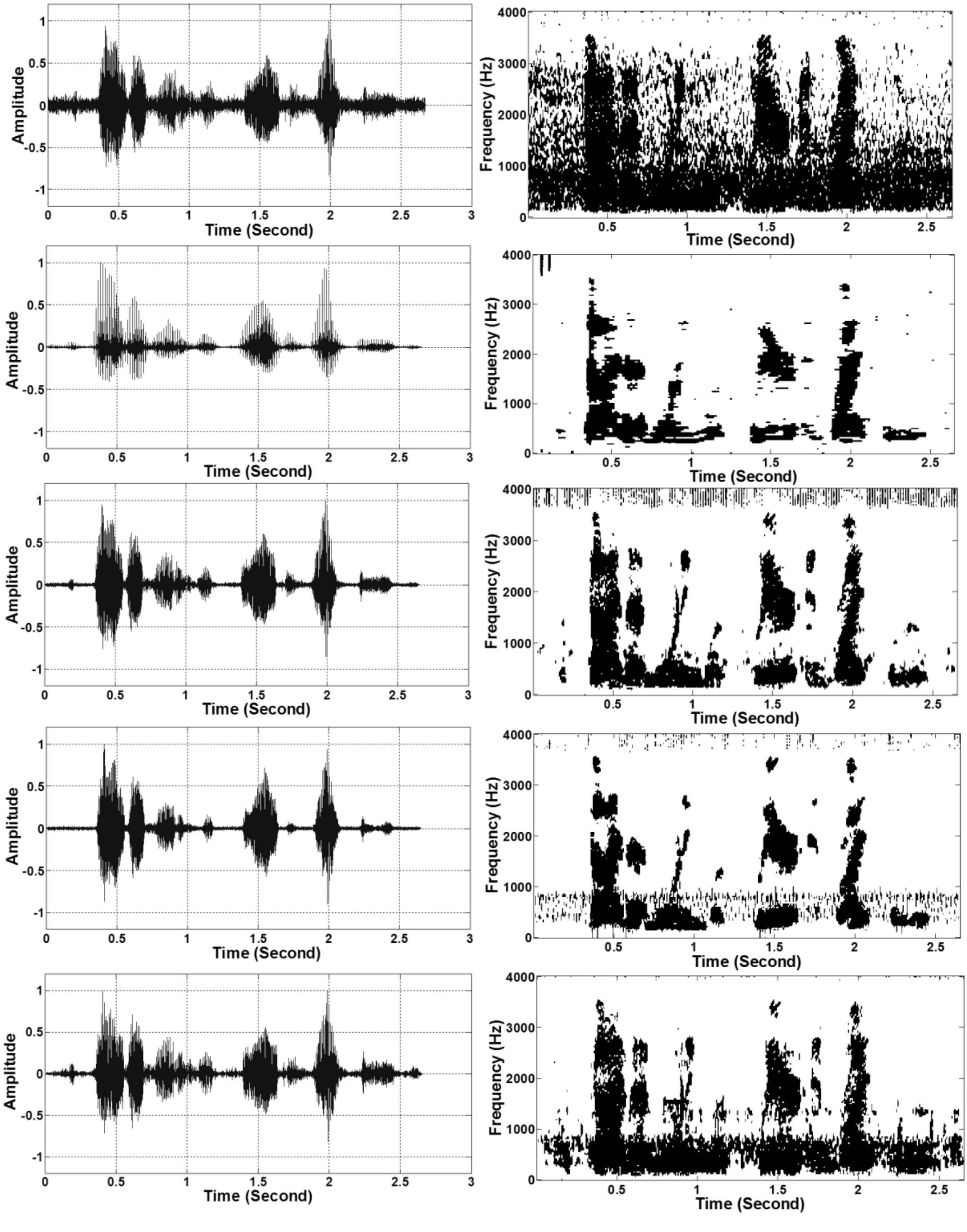


Fig. 6.12: Temporal waveforms and speech spectrogram of *sp10.wav* utterance, "The sky that morning was clear and bright blue," by a male speaker from the NOIZEUS corpus (From top to bottom) : (a) noisy speech (speech degraded by car noise at 10 dB SNR) (PESQ = 2.169); (b) speech enhanced by BSS algorithm (PESQ = 2.154); (c) speech enhanced by SOS algorithm (PESQ = 2.579); (d) speech enhanced by MBSS (PESQ = 2.259); and (e) speech enhanced by PMS-MBSS (PESQ = 2.426).

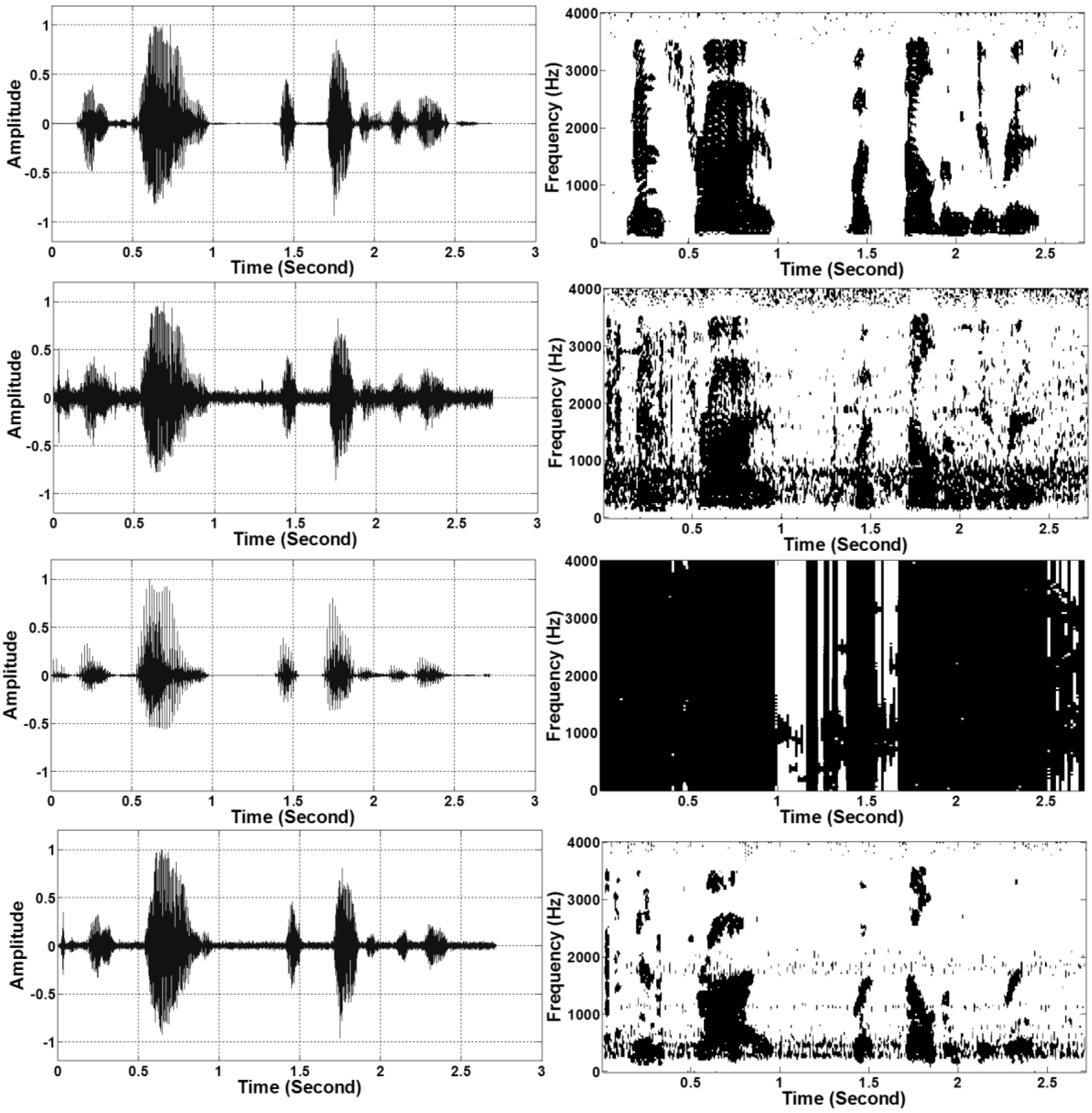


Fig. 6.13 (I): Temporal waveforms and speech spectrogram of *sp 6.wav* utterance, "Men strive but seldom get rich", by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech (PESQ = 4.5); (b) noisy speech (speech degraded by car noise at 10 dB SNR) (PESQ = 2.205); (c) speech enhanced by BSS algorithm (PESQ = 2.231); (d) speech enhanced by SOS algorithm (PESQ = 2.352).

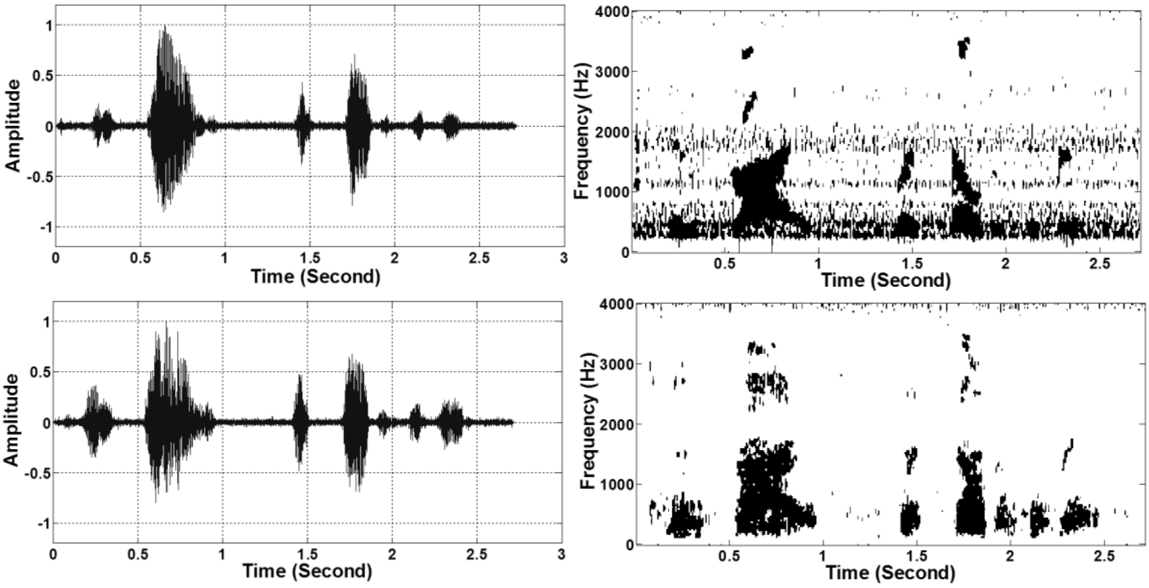


Fig. 6.13 (II): Temporal waveforms and speech spectrogram of *sp 6.wav* utterance, "Men strive but seldom get rich", by a male speaker from the NOIZEUS corpus (From top to bottom): (e) speech enhanced by MBSS algorithm (PESQ = 2.157); and (f) speech enhanced by PMS-MBSS (PESQ = 2.195).

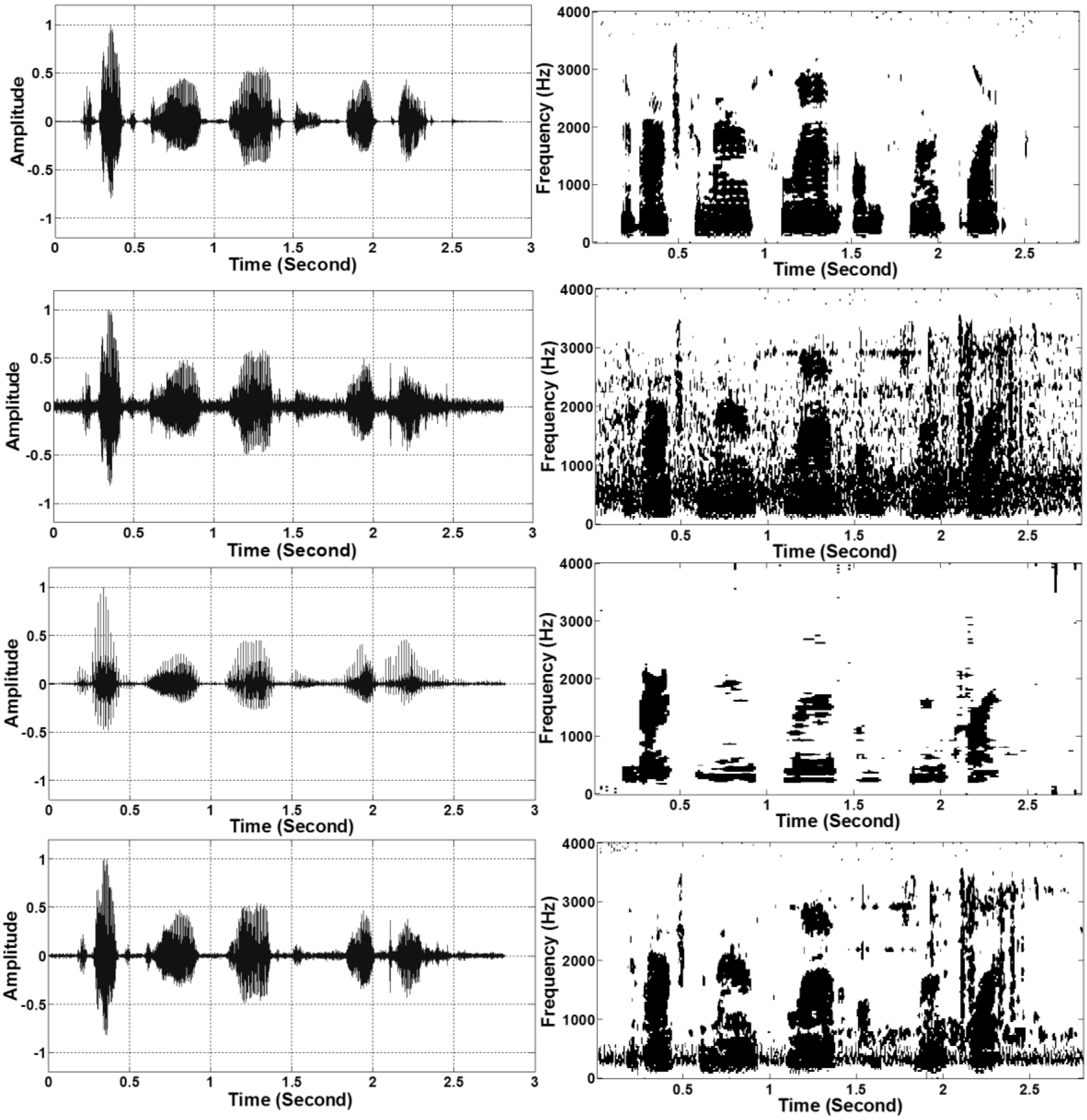


Fig. 6.14 (I) : Temporal waveforms and speech spectrogram of *sp1.wav* utterance, "The birch canoe slid on the smooth planks", by a male speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech (PESQ = 4.5); (b) noisy speech (speech degraded by car noise at 10 dB SNR) (PESQ = 2.084); (c) speech enhanced by BSS algorithm (PESQ = 1.898); (d) speech enhanced by SOS algorithm (PESQ = 2.167).

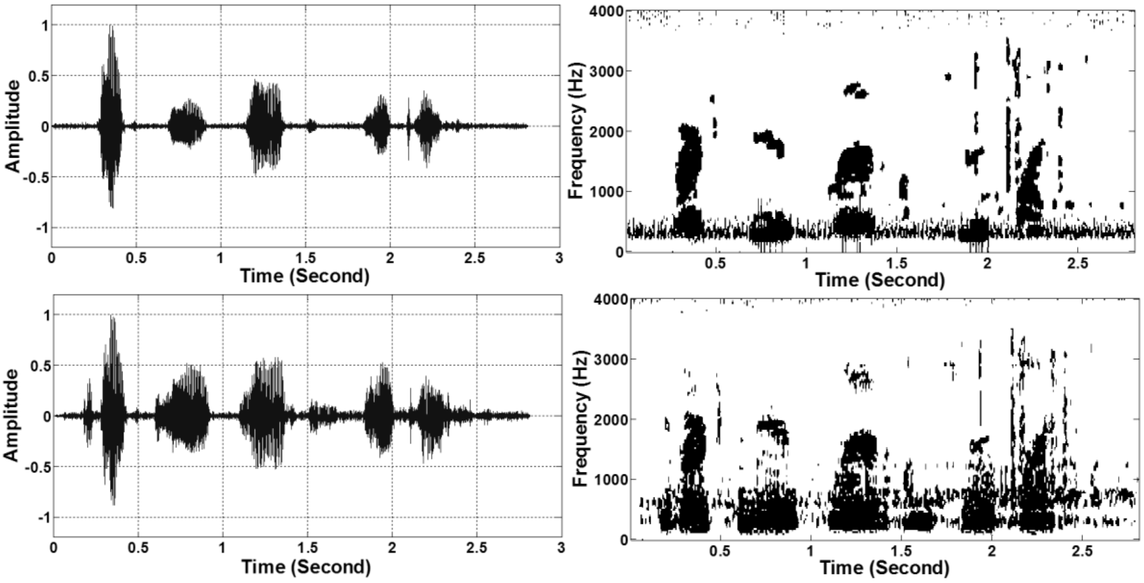


Fig. 6.14 (II): Temporal waveforms and speech spectrogram of *sp1.wav* utterance, "The birch canoe slid on the smooth planks", by a male speaker from the NOIZEUS corpus (From top to bottom): (e) speech enhanced by MBSS algorithm (PESQ = 2.030); and (f) speech enhanced by PMS-MBSS (PESQ = 2.276).

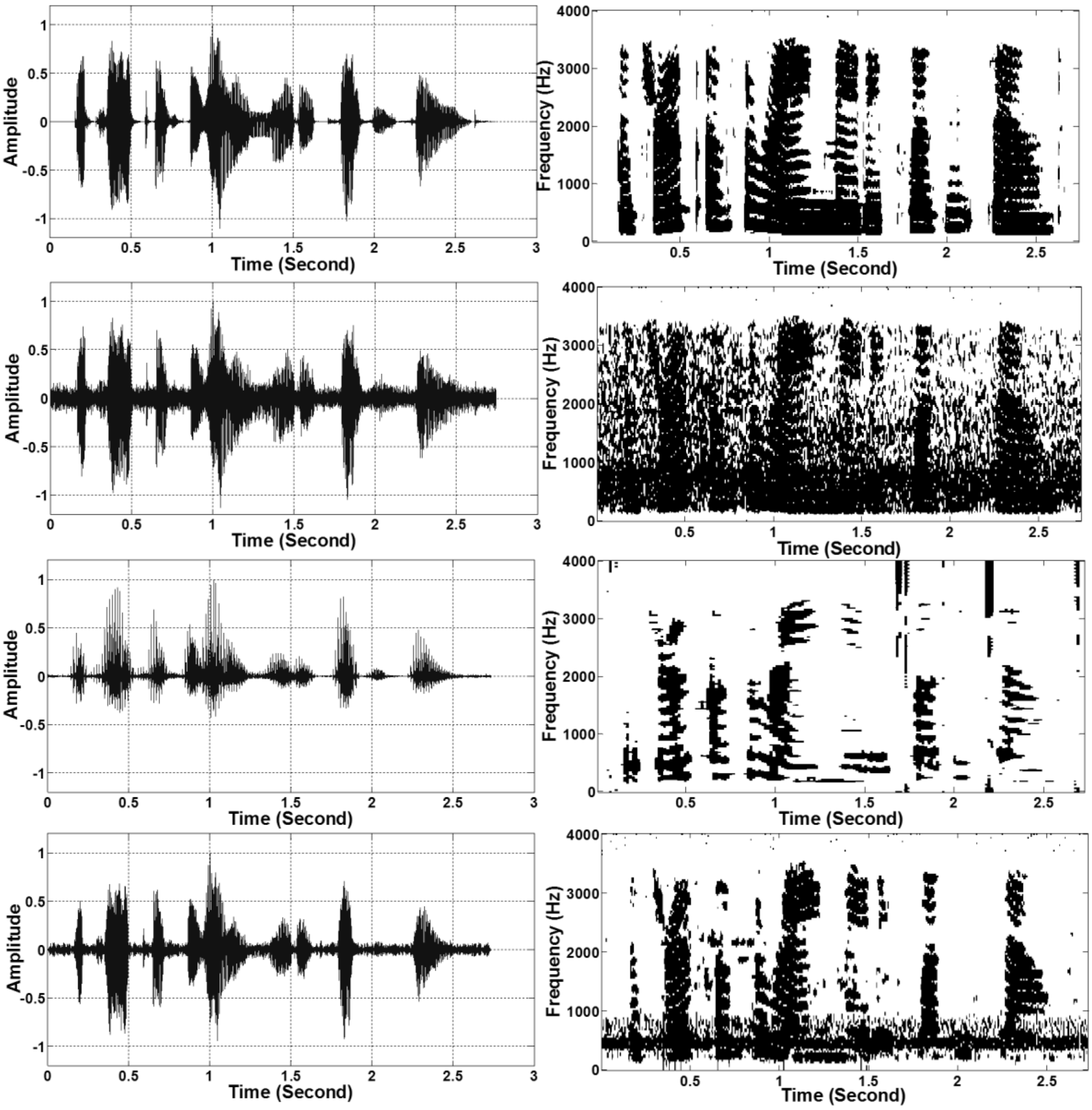


Fig. 6.15 (I): Temporal waveforms and speech spectrogram of *sp12.wav* utterance, "The drip of the rain made a pleasant sound", by a female speaker from the NOIZEUS corpus (From top to bottom): (a) clean speech (PESQ = 4.5); (b) noisy speech (degraded by car noise at 10 dB SNR) (PESQ = 2.043); (c) speech enhanced by BSS algorithm (PESQ = 1.782); (d) speech enhanced by SOS algorithm (PESQ = 2.341).

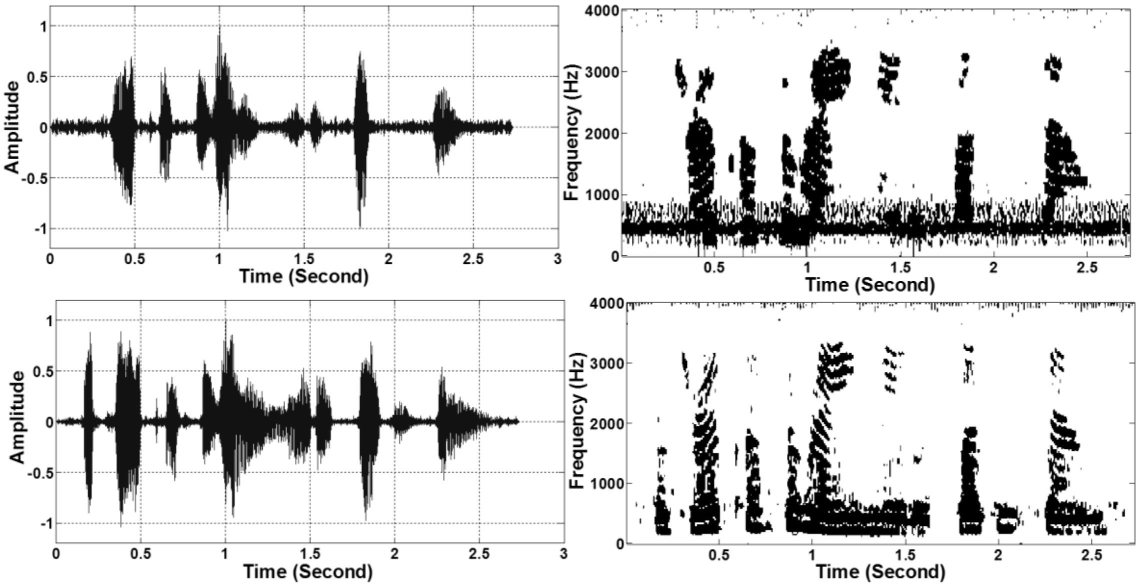


Fig. 6.15 (II): Temporal waveforms and speech spectrogram of *sp12.wav* utterance, "The drip of the rain made a pleasant sound", by a female speaker from the NOIZEUS corpus (From top to bottom): (e) speech enhanced by MBSS algorithm (PESQ = 2.005); and (f) speech enhanced by PMS-MBSS (PESQ = 2.242).

6.7. Summary

In this chapter, we have presented a speech enhancement algorithm which incorporates a perceptually motivated stationary wavelet packet filterbank (PM-SWPFB) (based on human auditory system) and an improved spectral over-subtraction (I-SOS) algorithm. The PM-SWPFB decomposes the input speech signal in the non-uniform subbands as per critical band rate scale and then spectral over-subtraction is applied in each subband, separately. The I-SOS uses an adaptive approach to estimate the noise from each subband. The adaptive noise estimation technique does not require explicit voice activity detection, thus providing accurate results even at very low SNR. The approach works continuously, even in the presence of speech. The selection of the appropriate wavelet filter for the SWPT ensures adequate frequency resolution for designing the PM-SWPFB.

The performance assessment of the proposed algorithm is done based on the various criteria such as, spectrogram plots, objective measures, and subjective listening tests. It is observed that the PMS-MBSS algorithm not only greatly reduces the noise but also prevents the speech spectrum getting deteriorated, especially at the low SNR noise cases. The superior results can be noticed for all types of noises of both non-stationary and stationary type at SNR (≤ 15 dB). This presented algorithms shows better performance than multi-band spectral subtraction and basic spectral subtraction algorithm at all levels of SNRs and shows better quality improvement than spectral over-subtraction algorithm at SNR (> 5 dB).

Moreover, results show that the speech quality obtained by using seventeen non-overlapping non-uniformly spaced frequency bands PMS-MBSS algorithm is better than the auditory perception based improved multi-band spectral subtraction (API-MBSS) algorithm with six non-overlapping non-uniformly spaced frequency bands, as described in Chapter 5. Further, both of these algorithms outperform the than an iterative processing based multi-band spectral subtraction (IP-MBSS) algorithm.

Chapter 7

Conclusion

7.1. Summary of Developments and Achievements

This chapter summarizes the body of work contained in this thesis with a gist of contributions along with the major results obtained from the research work conducted. Section 7.2 gives some indications of possible directions for future research.

This thesis has dealt with enhancement of single channel narrowband speech in adverse noisy conditions. The basic assumption about the noisy environments is that the additive background noise of both stationary and non-stationary types are additive in additive nature. The thesis has presented a set of transform based multi-band speech enhancement algorithms with two classes of multi-band criteria of both uniformly and non-uniformly spaced frequency bands. The uniformly spaced frequency bands based speech enhancement algorithm is proposed that utilizes iterative processing. Under the class of non-uniformly spaced frequency bands, two algorithms have been presented, namely, an improved multi-band spectral subtraction based on the critical band rate scale and a perceptually motivated stationary wavelet packet transform based improved spectral over-subtraction. These transform based multi-band speech enhancement algorithms are involved in identification and suppression of degraded speech regions and subsequent enhancement of these low SNR regions in frequency domain. Both the stationary and non-stationary types of noises have been considered in our work and the proposed three algorithms have been evaluated with extensive objective measures and subjective listening tests.

A family of spectral subtractive-type algorithms have been described and compared in an unified framework. This class of algorithms has been found to be attractive because of its simple implementation and ease of computation. Its major drawback is the introduction of un-natural remnant sound tones with musical structure in the enhanced speech. This noise can be more annoying to a human listener than the original background noise at very low SNRs. In addition to the basic magnitude spectral subtraction algorithm, several variations have been developed. Based on the definition of a gain function in the frequency domain, a unified formulation of these algorithms has been studied which multiplies the spectral magnitudes of the noisy speech signal to obtain the enhanced spectral magnitude. Next, a detailed study on stationary wavelet packet transform has been described. These works form the basis for the development of the three multi-band speech enhancement algorithms presented subsequently, in the thesis.

The first algorithm proposed in the thesis is the iterative processing based multi-band spectral subtraction algorithm for enhancement of degraded speech. In this proposed algorithm, the output of basic multi-band spectral subtraction (MBSS) algorithm has been used iteratively for progressively reducing the remnant noise, which is re-estimated in each iteration separately, in each band. The simulation results as well as subjective evaluations confirm that the speech enhanced by proposed algorithm is more pleasant to listeners than speech enhanced by conventional MBSS algorithms. From extensive simulation results and subjective listening tests it is observed that our iterative processing based multi-band spectral subtraction algorithm performs better than the classical MBSS and BSS for each type of noise (used in this work) at low SNRs (SNR < 5 dB).

An improved multi-band spectral subtraction based on critical band rate scale of human auditory system is presented next. In this algorithm, the bands are non-uniformly frequency spaced and noise estimate is updated by adaptively smoothing the noisy signal power. The simulation results as well as subjective evaluations show that the proposed algorithm reduces remnant noise efficiently and the enhanced speech contains minimal speech distortions (if any) with objective evaluation results. The noise estimation method used in our algorithm does not require the explicit voice activity detection and the noise estimate is updated adaptively and continuously even during speech activity from each band. The various objective measures and PESQ score suggest that performance of our algorithm shows better results at higher levels of SNR (> 5dB) except for the case restaurant and white Gaussian noise, in comparison to MBSS, BSS and SOS.

The last algorithm proposed in the thesis is the perceptually motivated stationary WPT based improved multi-band spectral over-subtraction speech enhancement algorithm. The stationary wavelet packet filterbank is obtained by adjusting the uniformly spaced stationary wavelet packet tree by temporally processing the input noisy speech signal in such a way it mimics closely with the critical bands structure of human ear. The noise estimate is updated by adaptively smoothing the noisy signal power independently, in the subbands, without the need of detecting speech pauses. It is inferred from the extensive simulation results and subjective listening tests that performance of the algorithm is superior to MBSS for SNR (> 0 dB), and BSS for each type of noise (that is considered in this thesis). In comparison to spectral over-subtraction algorithm, improvement in PESQ score is observed for car noise, train noise, and exhibition noise, at 0 dB SNR whereas poorer results are obtained for other type of noises.

From the study of speech spectrogram of the enhanced speech, as well as from the subjective listening tests it is confirmed that the remnant noise is less annoying to a human listener because of the modification of its musical structure to a perceptually white quality while keeping the speech distortions in an acceptable limit. The performance measure leads to the conclusion that the overall SNR improvement does not provide a good indicator to speech quality, whereas objective measures such as SegSNR, Itakura-Saito distortion (ISD), and PESQ score provide a much better indicator for evaluation of speech quality.

The performance analysis study of the proposed algorithms lead us to conclude that the first proposed algorithm, i.e., iterative processing based multi-band spectral subtraction perform mostly well for all types of noises of both non-stationary and stationary types at low SNRs. It is also observed that this algorithm does not work well for high SNR (> 5 dB). The concept of iterative processing is found to be effective for uniformly spaced multi-band case. In case of non-uniform multi-bands, the effect of iterative processing does not seem to have much effect.

The next algorithm proposed in this thesis, i.e., critical band rate sale based improved multi-band spectral subtraction, shows superior results for most types noises (non-stationary and stationary) at low SNR (≤ 10 dB) except restaurant noise, while for SNR (≈ 15 dB) it shows better results for car, exhibition, white (stationary) noises and gives poorer results for street noise. This algorithm shows better performance than multi-band spectral subtraction and basic spectral subtraction algorithm at all levels of SNRs; also, shows comparable performance with spectral over-subtraction algorithm except for street noise case. The optimum performance of this STFT based algorithm with adaptive noise estimation is achieved with six non-uniformly spaced frequency bands, matching closely with the critical band rate scale of human ear.

The perceptually motivated stationary wavelet packet transform based improved multi-band spectral over-subtraction algorithm shows superior results for all types noises (non-stationary and stationary) at all levels of SNR (≤ 15 dB). This algorithm shows better performance than multi-band spectral subtraction and basic spectral subtraction algorithm at all levels of SNRs; also, shows better performance than spectral over-subtraction algorithm at SNR (> 5 dB). This algorithm utilizing band-specific speech enhancement, adaptive noise estimation and the auditory frequency scale inspired subbands, gives the best performance with respect to the other two presented algorithms.

Thus, the proposed algorithms have shown very promising results in various noisy conditions. Compared to other spectral subtractive-type algorithms with uniformly spaced frequency bands, and with fixed value of subtractive parameters, the speech enhancement obtained from the proposed algorithms is reasonably good. The adaptation of the subtraction parameters in the proposed algorithms leads to a significant reduction of the unnatural structure of musical remnant noise. The introduction of iterative processing and auditory perception criterion (critical band rate scale) in transform domain as well as in temporal-transform domain allows us to obtain remnant noise with perceptually white quality in various noisy environments and even at very low SNRs (SNR < 5 dB).

We can summarize that, the transform based multi-band speech enhancement algorithms provide a definite improvement over the spectral subtractive-type algorithms and do not suffer from musical remnant noise to a great extent. The improvement in speech quality and intelligibility can be attributed to the multi-band approach, processing of signal in accordance to the human auditory system, the iterative processing and the non-uniform effect of non-stationary noise on the spectrum of speech.

7.2. Scope for Future Work

Although many issues of transform based multi-band speech enhancement algorithms have been investigated, there are still several important topics that provide focus on future research. Some of the possible research directions are depicted below:

- There is a need to develop more robust and accurate voice activity detection (VAD) algorithm which preserves the transitional regions and unvoiced regions of speech signal containing low energy levels. This will help in improving the speech enhancement algorithms performance in terms of quality and intelligibility of enhanced speech.
- The empirically derived values for the additional band subtraction parameters have been used for the presented algorithms in the thesis. An automated way of adaptively, calculating the suitable value of additional band subtraction factor in place of empirically derived value will be an exciting direction of future work.
- Voiced speech which contains higher energy segments seem to be more important than un-voiced speech for preserving speech quality. Therefore, besides suppressing the background noise, algorithms may be developed that boost up the voiced part from the speech signal.

- An alternative approach to calculate the over-subtraction factors in each band/subband can be investigated as the use of the *logarithmic* function is computationally expensive for implementation in real-time systems.
- For the stationary WPT based speech enhancement algorithm, there is an inherent time delay which exists irrespective of implementation method one uses, for the decomposition of the noisy speech. There is a trade-off between the performance and delay time while selecting the wavelet filter. Generally, the longer wavelet filter has a better frequency response with a longer time delay whereas the shorter wavelet filter has a poorer frequency response with a shorter time delay. Therefore, obtaining a mother wavelet which gives superior frequency response with shorter support will be more desirable.
- Fixed-point digital signal processors (DSPs) are becoming increasingly popular in applications such as mobile phones, and digital hearing aids due to their low-power consumption and high processing rate. The transform based multi-band speech enhancement algorithms, as presented in this thesis can be implemented in real-time on a fixed-point digital signal processor platform in real-world conditions.

Appendix A

Overlap-Addition Method

To construct a signal, $y(n)$, which ideally should be the same as the original signal, $s(n)$, the overlap-add (OLA) method requires that the inverse discrete Fourier transform (DFT) be taken for each frame in the discrete STFT. Each of these short-time sections are then overlapped and added [3, 19, 28]:

$$y(n) = \frac{1}{W(0)} \sum_{p=-\infty}^{\infty} \left[\frac{1}{N} \sum_{k=0}^{N-1} S(p,k).\, e^{j2\pi nk/N} \right) \Big] \tag{A.1}$$

where

$$W(0) = \sum_{n=-\infty}^{\infty} w(n) \tag{A.2}$$

We can rewrite (A.1) as

$$y(n) = \frac{1}{W(0)} \sum_{p=-\infty}^{\infty} s(n) w(p-n) = \frac{1}{W(0)} s(n) \sum_{p=-\infty}^{\infty} w(p-n) \tag{A.3}$$

Therefore, for $y(n) = s(n)$, the following constraint must be met:

$$\sum_{p=-\infty}^{\infty} w(p-n) = W(0) \tag{A.4}$$

This is the OLA constraint. This constraint requires that the sum of the analysis windows (which are obtained by sliding $w(n)$ by 1 time sample) add up to the same value at each discrete point in time. This results in the elimination of the analysis window from the synthesized sequence.

Further, if the STFT is decimated by factor L, then,

$$y(n) = \frac{L}{W(0)} \sum_{p=-\infty}^{\infty} \left[\frac{1}{N} \sum_{k=0}^{N-1} S(pL,k).\, e^{j2\pi nk/N} \right) \Big] \tag{A.5}$$

We can rewrite (A.5) as

$$y(n) = \frac{L}{W(0)} \sum_{p=-\infty}^{\infty} s(n)w(pL-n) = \frac{L}{W(0)} s(n) \sum_{p=-\infty}^{\infty} w(pL-n) \tag{A.6}$$

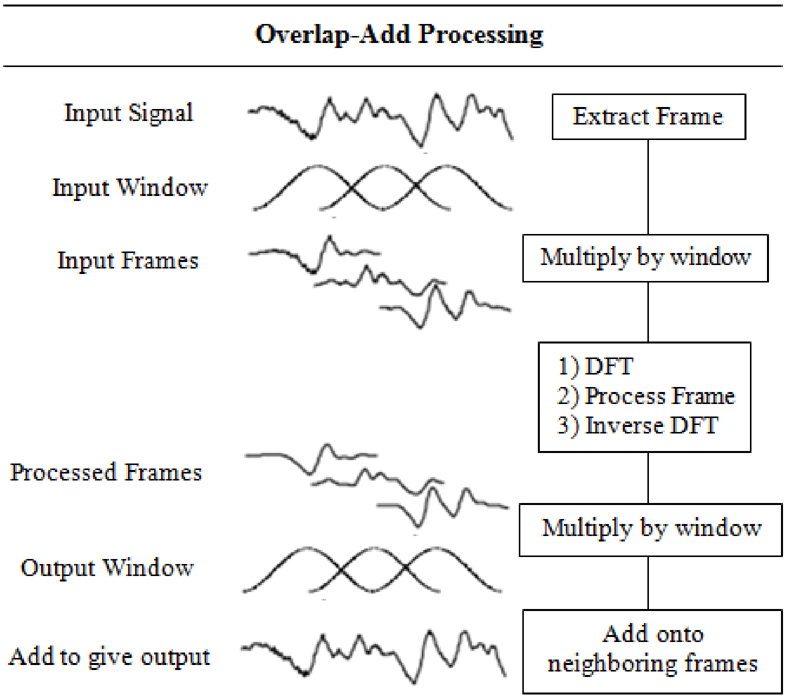
Therefore, for $y(n) = s(n)$, the following constraint must be met:

$$\sum_{p=-\infty}^{\infty} w(pL-n) = \frac{W(0)}{L} \tag{A.7}$$

This is the generalised OLA constraint. This constraint requires that the sum of the analysis windows (which are obtained by sliding $w(n)$ by L time samples) add up to the same value at each discrete point in time. This process is graphically depicted in Table A.1.

Table A.1

Overlap-Add Processing



Appendix B

MOS Test Listeners' Profile

The names and email addresses of the listeners who participated in the MOS test are given in the following table:

TABLE B.1 MOS TEST LISTENERS' PROFILE

Initial	**Name**	**Sex**	**Email**
VLP	Lt.Col. Vikrant Lakhan Pal	Male	vikrantlakhandipal@rediffmail.com
JCJ	Lt. Col. J.C. Joshi	Male	jcjoshi@gmail.com
RM	Roshan Mathew	Male	mails4roshan@gmail.com
AN	Ashwin N.	Male	ashwin.nenmini@gmail.com
AMS	Abhishek M.S.	Male	msabhishek.ms@gmail.com

References

[1]. D. O'Shaughnessy, *Speech Communications: Human and Machine*, IInd ed. Hyderabad, India: University Press (India) Pvt. Ltd., 2007.

[2]. Y. Ephraim, "Statistical-model-based speech enhancement systems," in *Proceedings of the IEEE*, Oct. 1992, vol. 80, no. 10, pp. 1526–1555.

[3]. P. C. Loizou, *Speech Enhancement: Theory and Practice*, Boca Raton, FL, USA: CRC Press, Taylor & Francis Group, 2007.

[4]. J. S. Lim, *Speech Enhancement*, Englewood Cliffs, NJ: Prentice–Hall, 1983.

[5]. L. Rebiner, and B. H. Juang, *Fundamentals of Speech Recognition*, Upper Saddle River, NJ, USA: Prentice-Hall PTR, 1993.

[6]. John C. Ballamy, *Digital Telephony*, IIIrd ed., Wiley Publication, 2000.

[7]. B. H. Juang, "Recent developments in speech reorganization under adverse conditions," in *Proceedings of International Conference on Spoken Language Processing*, Japan, Nov. 18–22, 1990, pp. 1113–1116.

[8]. Yifan Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, April 1995.

[9]. Y. Ephraim, H. L. Ari, and W. Roberts, "A brief survey of speech enhancement," in *the Electrical Engineering Handbook*, IIIrd ed. Boca Raton, FL: CRC press, 2006.

[10]. Y. Ephraim, and I. Cohen, "Recent advancements in speech enhancement," in *the Electrical Engineering Handbook*, CRC press, ch. 5, pp. 12–26, 2006.

[11]. J. R. Deller, J. G. Prokis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals*, Ist ed. Upper Saddle River, NJ, USA: Prentice-Hall PTR, 1993.

[12]. L. B. Thomas, and A. Ravindran, "Intelligibility enhancement of already noisy speech signals," *Journal of the Audio Engineering Society*, vol. 22, no. 4 , pp. 234–236, 1974.

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

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

[15]. Ning Ma, M. Bouchard, and R. A. Goubran, "Perceptual Kalman filtering for speech enhancement in colored noise," in *Proceedings of IEEE Acoustics, Speech, and Signal Processing,* Montreal, Canada, May 17–21, 2004, vol. 1, pp. 717–720.

[16]. Ning Ma, M. Bouchard, and R. A. Goubran, "A wavelet Kalman filter with perceptual masking for speech enhancement in colored noise ," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2005, vol. 1, pp. 149–152.

[17]. Z. Goh, K. C. Tan, and B. T. G. Tan, "Kalman filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 7, no. 5, pp. 510–524, Sept.1999.

[18]. Y. Eprahim, and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 5, pp. 1109–1121, Dec. 1984.

[19]. Leigh D. Alsteris, and Kuldip K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing,* vol. 17, no. 3, pp. 578–616, May 2007.

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

[21]. S. Kamath, and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, USA, May 13-17, 2002, vol. 4, pp. 4160–4164.

[22]. M. A. Abd El-Fattah, M. I. Dessouky, S. M. Diab, and F. E. Abd El-samie, "Speech enhancement using an adaptive Wiener filtering approach," *Progress In Electromagnetics Research M.*, vol. 4, pp. 167–184, 2008.

[23]. S. Ogata, and T. Shimamura, "Reinforced spectral subtraction method to enhance speech signal," in *Proceedings of IEEE International Conference on Electrical and Electronic Technology*, 2001, vol. 1, pp. 242–245.

[24]. N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech, and Audio Processing*, vol. 7, no. 2 , pp. 126–137, March 1999.

[25]. S. Mallat, *A Wavelet Tour of Signal Processing*, IInd ed., San Diego, Academic-Press, 1999.

[26]. Raghuveer M. Rao, and Ajit S. Bopardikar, *Wavelet Transforms: Introduction to Theory and Applications*, Wellesley-Cambridge Press, 1998.

[27]. D. L. Donoho, "Denoising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[28]. L. R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Ist ed. Pearson Education, 1978.

[29]. S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, IInd ed. NY, USA: Wiley, 2000.

[30]. S. H. Chen, and J. F. Wang, "Speech enhancement using perceptual wavelet packet decomposition and teager energy operator," *Journal of VLSI Signal Processing*, vol. 36, no. 2–3, pp. 125–139, March 2004.

[31]. Y. Ghanbari, and M. R. Karami Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets," *Speech Communication*, vol. 48, no. 8, pp. 927–940, Aug. 2006.

[32]. Yu Shao, and Chip-Hong Chang, "A generalized time-frequency subtraction method for robust speech enhancement based on wavelet filterbanks modeling of human auditory system," *IEEE Transactions on Systems, Man, and Cybernetic-Part B*: *Cybernetics*, vol. 37, no. 4, pp. 877–888, Aug. 2007.

[33]. C. T. Lu, and H. C. Wang, "Enhancement of single channel speech based on masking property and wavelet transform," *Speech Communication*, vol. 41, no. 2–3 , pp. 409–427, Oct. 2003.

[34]. C. T. Lu, and H. C. Wang, "Speech enhancement using perceptually constrained gain factors in critical band wavelet packet transform," *Electronics Letters*, vol. 40, no. 6, pp. 394–396, March 2004.

[35]. C. T. Lu, and H. C. Wang, "Speech enhancement using hybrid gain factor in critical band wavelet packet transform," *Digital Signal Processing*, vol. 17, no. 1, pp. 172–178, Jan. 2007.

[36]. T. Glzow, A. Englesberg, and U. Heute, "Comparison of discrete wavelet transformation and a non-uniform polyphase filterbank applied to spectral subtraction speech enhancement," *Signal Processing*, vol. 64, no. 1, pp. 5–19, Jan. 1998.

[37]. A. Papoulis, and S. Pillai, *Probability Random Variables and Stochastic Processes*, IVth ed. NY: McGraw-Hill, 2002.

[38]. J. S. Lim, and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," in *Proceedings of the IEEE*, Dec. 1979, vol. 67, no. 12, pp. 1586–1604.

[39]. L. W. David, and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, Aug. 1982.

[40]. R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis and Synthesis," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 28, pp. 99–102, 1980.

[41]. P. Lockwood, and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and projection, for robust recognition in cars," *Speech Communication,* vol. 11, no. 2–3, pp. 215–228, 1992.

[42]. P. Handel, "Low-distortion spectral subtraction for speech enhancement," in *Proceedings of European Conference on Speech Communication and Technology,* Madrid, Spain, Sept. 18–21, 1995, vol. 2, pp. 1549–1552.

[43]. L. Arslan, A. McCree, and V. Vishwanathan, "New methods for adaptive noise suppression," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Proceesing*, Dallas, USA , May 9–12, 1995, vol. 1, pp. 812–815.

[44]. H. G. Hirsh, and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Proceesing*, Dallas, USA, May 9–12, 1995, vol. 1, pp. 153–156.

[45]. R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of European Signal Processing Conference,* Edinburg, U.K., Sept. 1994, pp. 1182–1185.

[46]. G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proceedings of Euro Speech*, 1995, vol. 2, pp 1513–1516.

[47]. Sheng Li, Jian Qi Wang, and Xi Jing Jing, "The application of non-linear spectral subtraction method on millimeter wave conducted speech enhancement," *Mathematical Problems in Engineering,* pp. 1–12, 2010.

[48]. R. M. Udrea, N. Vizireanu, S. Ciochina, and S.Halunga, "Non-linear spectral subtraction method for colored noise reduction using multi-band Bark scale," *Signal Processing*, vol. 88, no. 5, pp. 1299–1303, 2008.

[49]. Venkata Rama Rao, Rama Murthy, and K. Srinivasa Rao, "Speech enhancement using cross-correlation compensated multi-band Wiener filter combined with harmonic regeneration," *Journal of Signal and Information Processing*, vol. 2, no. 2, pp. 117–124, May 2011.

[50]. Y. Ghanbari, M. R. K. Mollaei, and B. Amelifard, "Improved multi-band spectral subtraction method for speech enhancement," in *Proceedings of International Conference on Signal, and Image Processing*, Hawaii, USA, Aug. 23–25, 2004, pp. 225–230.

[51]. Zenton Goh, Kah-Chye Tan, and B. T. G. Tan, "Post-processing method for suppressing musical noise generated by spectral subtraction," *IEEE Transactions on Speech, and Audio Processing*, vol. 6, no. 3, pp. 287–292, May 1998.

[52]. K. Yamashita, S. Ogata, and T. Shimamura, "Improved spectral subtraction utilizing iterative processing," *Electronics and Communications*, Japan, vol. 90, no. 4, pp. 39–51, 2007.

[53]. K. Yamashita, S. Ogata, and T. Shimamura, "Spectral subtraction iterated with weighting factors," in *Proceedings of IEEE Speech Coding Workshop*, Oct. 6–9, 2002, pp.138–140.

[54]. Sheng Li, Jian-Qi Wang, Ming Niu, Xi-Jing Jing, and Tian Liu, "Iterative spectral subtraction method for millimeter wave conducted speech enhancement," *Journal of Biomedical Science and Engineering*, vol. 3, no. 2, pp. 187–192, Feb. 2010.

[55]. R. M. Udrea, N. D. Vizireanu, and S. Ciochina, "An improved spectral subtraction method for speech enhancement using perceptual weighting filter," *Digital Signal Processing*, vol.18, no. 4, pp.581–587, July 2008.

[56]. James D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal of Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[57]. S. Mallat, "A theory for multi-resolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, July 1989.

[58]. C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms*: *A Primer*, Prentice-Hall, New Jersey, 1998.

[59]. W. J. Phillips, "Wavelet and Filterbanks Course Notes," 2003. [Available Online] http://www.engmath.dal.ca/courses/engm6610/notes/notes.html

[60]. G. Strang, and T. Nguyen, *Wavelets and Filterbanks*. Wellesley, MA: Wellesley-Cambridge Press, IInd ed., 1997.

[61]. R. R. Coifman, and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713–718, March 1992.

[62]. S. Olhede, and Andrew T. Walden, *"*A generalized demodulation approach to time-frequency projections for multi-component signals,*"* in *Proceedings of the Royal Society London- Series A*, 2005, vol. 461, pp. 2159–2179.

[63]. Andrew T. Walden, and Alberto Contreras Cristan, "The phase-corrected un-decimated discrete wavelet packet transform and its application to interpreting the timing of events," in *Proceedings of the Royal Society London-Series A*, 1998, vol. 454, no. 1976, pp. 2243–2266.

[64]. I. Cohen, S. Raz, and David Malah, "Orthonormal shift-invariant wavelet packet decomposition and representation," *Signal Processing*, vol. 57, no. 3, pp. 251–270, 1997.

[65]. Andrew T. Walden, and Alberto Contreras Cristan, "The phase-corrected un-decimated wavelet packet transform and the recurrence of high latitude interplanetary shock waves," in *Proceedings of the Royal Society London-Series A*, Aug. 1997, pp. 0–26.

[66]. I. Cohen, "Enhancement of speech using Bark scaled wavelet package decomposition," in *Proceedings of EUROSPEECH*, Denmark, Sept. 3–7, 2001.

[67]. H. Tasmaz, and E. Ercelebi, "Speech enhancement based on un-decimated wavelet packet perceptual filterbanks and MMSE-STSA estimation in various noise environments," *Digital Signal processing*, vol. 18, no. 5, pp. 797–812, Sept. 2008.

[68]. M. S. Koh, and M. Mortz, "Speech enhancement based on truncated an constrained minimum variance estimator (TCMVE) and un-decimated wavelet packet non-uniform filterbanks," in *Proceedings of Asilomar Conference on Signals, System and Computers*, Monterrey, CA, 2001, vo. 1, pp. 538–544.

[69]. Y. Hu, and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7, pp. 588–601, 2007.

[70]. Y. Hu, and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no.1, pp. 229–238, 2008.

[71]. H. J. M. Steeneken, "Subjective and objective intelligibility measures," in *Proceedings of ESCA Workshop Speech Processing in Adverse Conditions*, Cannes, France, Nov. 1992, pp.1–10.

[72]. J. H. L. Hansen, and D. A. Cairns, "ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environment," *Speech Communication*, vol. 16, no. 4, pp. 391–422, June 1995.

[73]. O. Cappe', and J. Laroche, "Evaluation of short-time spectral attenuation techniques for the restoration of musical recording," *IEEE Transactions on Speech, and Audio Processing*, vol. 3, no. 1, pp. 84–93, Jan. 1995.

[74]. A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, 2001, vol. 2, pp. 749–752.

[75]. "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Recommendation P.862, Feb. 2001.

[76]. A Noisy Speech Corpus for Evaluation of Speech Enhancement Algorithms. http://www.utdallas.edu/~loizou/speech/noizeus/

[77]. S. Nandkumar, and J. H. L. Hansen, "Dual-channel iterative speech enhancement with constraints on an auditory-based spectrum," *IEEE Transactions on Speech, and Audio Processing*, vol. 3, no. 1, pp. 22–34, Jan. 1995.

[78]. R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech, and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[79]. I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech, and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[80]. I. Cohen, and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[81]. I. Cohen, "Speech enhancement using a non-causal a priori SNR Estimator," *IEEE Signal Processing Letters,* vol. 11, no. 9, pp. 725–728, Sept. 2004.

[82]. L. Lin, W. H. Holmes, and E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement," *Electronics Letters,* vol. 39, no. 9, pp. 754–755, May 2003.

[83]. L. Lin, W. Holmes, and E. Ambikairajah, "Speech denoising using perceptual modification of Wiener filtering," *Electronics Letters*, vol. 38, no. 23, pp. 1486–1487, Nov. 2002.

[84]. E. Zwicker, and H. Fastl, *Psychoacoustics: Facts and Models,* IInd ed., Germany, Springer-Verlag, Berlin Heidelberg, New York, 1990.

[85]. E. Zwicker, and E. Terhardt, "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," *Journal of the Acoustical Society of America*, vol. 68, no. 5, pp. 1523–1525, 1980.

[86]. L. Lin, W.H. Holmes, and E. Ambikairajah, "Subband noise estimation for speech enhancement using a perceptual wiener filter," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 6–10, 2003, vol. 1, pp. 80–83.

[87]. Peter L. Chu, and David G. Messerschmitt, "A frequency weighted Itakura-saito spectral distance measure," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 545–560, Aug. 1982.

[88]. Novlene Zoghlami, and Zied Lachiri, "Application of perceptual filtering models to noisy speech signals enhancement," *Journal of Electrical and Computer Engineering* pp. 1-12, 2012.

[89]. Arata Kawamura, Weerawut Thanhikam, and Youji Iiguni, "Single channel speech enhancement techniques in spectral domain," *ISRN Mechanical Engineering,* pp.1-9, 2012.

[90]. Kuldip Paliwal, Kamil Wo´jcicki, and Belinda Schwerin, "Single channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication,* vol. 52, no. 5, pp. 450–475, May 2010.

[91]. Kuldip Paliwal, Belinda Schwerin, and Kamil Wo´jcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Communication*, vol. 54, no. 4, pp. 282–305, 2012.

List of Publications

Journal Publications

I. Navneet Upadhyay, and Abhijit Karmakar, "Spectral subtractive-type algorithms for enhancement of noisy speech: an integrative review," *International Journal of Image, Graphics and Signal Processing*, 2013. [Accepted, To be published in June'2013]

II. Navneet Upadhyay, and Abhijit Karmakar, "A multi-band speech enhancement algorithm exploiting iterative processing for enhancement of single channel speech," *Journal of Signal and Information Processing*, vol. 4, no. 2, pp., 2013. [Under Print]

III. Navneet Upadhyay, and Abhijit Karmakar, "Single-channel speech enhancement using critical-band rate scale based improved multi-band spectral subtraction," *Journal of Signal and Information Processing*, vol. 4, no. 3, pp., 2013. [Under Print]

Conference Publications

I. Navneet Upadhyay, and Abhijit Karmakar, "A multi-band spectral subtraction using iterative processing for enhancement of degraded speech," in *Proceedings of International Conference on Signal, Image, and Video Processing*, IIT Patna, India, Jan. 13–15, 2012, pp. 192–196.

II. Navneet Upadhyay, and Abhijit Karmakar, "The spectral subtractive-type algorithms for enhancing speech in noisy environments," in *Proceedings of 1st IEEE International Conference on Recent Advances in Information Technology*, ISM Dhanbad, India, March 15 –17, 2012, pp. 841–847.

III. Navneet Upadhyay, and Abhijit Karmakar, "A perceptually motivated multi-band spectral subtraction algorithm for enhancement of degraded speech," in *Proceedings 3rd IEEE International Conference on Computer, and Communication Technology,* MNNIT Allahabad India, Nov. 23–25, 2012, pp. 340–345.

IV. Navneet Upadhyay, and Abhijit Karmakar, "Single channel speech enhancement utilizing iterative processing of multi-band spectral subtraction algorithm," in *Proceedings of 2nd*

IEEE International Conference on Power, Control, and Embedded System, MNNIT Allahabad, Dec. 17–19, 2012, pp. 196–201.

V. Navneet Upadhyay, and Abhijit Karmakar, "An auditory perception based improved multi-band spectral subtraction algorithm for enhancement of speech degraded by non-stationary noises," in *Proceedings of 4th IEEE International Conference on Intelligent Human Computer Interaction*, IIT Kharagpur, India, Dec. 27–29, 2012, pp. 392–398.

VI. Navneet Upadhyay, and Abhijit Karmakar, "A perceptually motivated stationary wavelet filter-bank utilizing improved spectral over-subtraction algorithm for enhancing speech in non-stationary environments," in *Proceedings of 4th IEEE International Conference on Intelligent Human Computer Interaction*, IIT Kharagpur, India, Dec. 27–29, 2012, pp. 472–478.

VII. Navneet Upadhyay, and Abhijit Karmakar, "A frequency dependent algorithm for enhancement of degraded speech," presented in National Conference on VLSI Design and Embedded Systems, CEERI Pilani, India, Oct. 12–14, 2011.

Manuscripts Submitted

I. Navneet Upadhyay, and Abhijit Karmakar, "A perceptually motivated stationary wavelet filterbank exploiting improved spectral over-subtraction for enhancement of degraded speech," in *Digital Signal Processing.*

II. Navneet Upadhyay, and Abhijit Karmakar, "Critical band rate scale based improved multi-band spectral subtraction for enhancement of speech in various noise environments," in *International Journal of Speech Technology.*

III. Navneet Upadhyay, and Abhijit Karmakar, "An improved multi-band spectral subtraction algorithm for enhancing speech degraded by non-stationary noises," in *International conference on Design and Manufacturing.*

VITAE

Navneet Upadhyay, received the Bachelor of Engineering degree in electronics and communication engineering discipline from Dr. B. R. Ambedkar University, Agra, India, in 2000, and the Master of Technology degree in digital communication discipline from Uttar Pradesh Technical University, Lucknow, India in 2006. Currently, he is pursuing his doctoral studies in speech signal processing area in electrical and electronics engineering department of Birla Institute of Technology & Science, Pilani, India.

His research interests are in the areas of speech processing (particularly speech enhancement, speech recognition, and speech coding) and digital communication.

Abhijit Karmakar was born in West Bengal, India, in 1971. He received the B.E. degree in electronics and telecommunication engineering from Jadavpur University, India, in 1993, the M.Tech degree in electrical engineering from Indian Institute of Technology Madras, Chennai, India, in 1995, and the Ph.D degree in electrical engineering from Indian Institute of Technology, Delhi, India, in 2007.

Since 1995, he is associated with Central Electronics Engineering Research Institute / Council of Scientific & Industrial Research (CSIR - CEERI), Pilani, India. His research interests are is the areas of digital signal processing, auditory modeling, speech quality evaluation, and VLSI design.