A Spectral Domain Speech Enhancement Method Based on Noise Compensations in Both Magnitude and Phase Spectra

by

Asaduzzaman

MASTER OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING

Department of Electrical and Electronic Engineering
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

July 2012

The thesis entitled **"A Spectral Domain Speech Enhancement Method Based on Noise Compensations in Both Magnitude and Phase Spectra"** submitted by Asaduzzaman, Student No.: 1009062094F, Session: October, 2009 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of MASTER OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING on July 11, 2012.

BOARD OF EXAMINERS

1. _____

(Dr. Celia Shahnaz) — **Chairman** (Supervisor)
Assistant Professor
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka - 1000, Bangladesh.

2. _____

(Dr. Pran Kanai Saha) — **Member** (Ex-officio)
Professor and Head
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka - 1000, Bangladesh.

3. _____

(Dr. Mohammed Imamul Hassan Bhuiyan) — **Member**
Associate Professor
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka - 1000, Bangladesh.

4. _____

(Dr. Farruk Ahmed) — **Member** (External)
Professor
School of Engineering and Computer Science (SECS)
Independent University, Bangladesh (IUB)
Block-B, Bashundhara R/A, Dhaka-1229.

CANDIDATE'S DECLARATION

I, do, hereby declare that neither this thesis nor any part of it has been submitted elsewhere for the award of any degree or diploma.

Signature of the Candidate

Asaduzzaman

Dedication

To my parents.

Acknowledgment

I would like to express my heartiest thanks and sincere gratitude to my supervisor Dr. Celia Shahnaz for her guidance, encouragement, constructive suggestions, and support during the span of this research. Among many things I learned from Dr. Shahnaz, endless effort for achieving excellence in research is undoubtedly the most important one, which enables me to understand and solve many important problems in the area of speech enhancement. I believe our work in the past one and half year advanced the state of art of the speech enhancement problem. I also want to thank her for spending her valuable time in exploring new ideas and improving the writing of this thesis.

I would also like to thank the rest of the members of my thesis committee: Prof. Dr. Md. Pran Kanai Saha, Dr. Mohammed Imamul Hassan Bhuiyan, and Prof. Dr. Farruk Ahmed, for their encouragement and insightful comments. I would like to thank the head of the department of Electrical and Electronic Engineering for allowing me to use the lab facilities, which contributed greatly in completing the work in time. I wish to express note of thanks to Dr. Shaikh Anowarul Fattah, for providing inspiration and academic guidance during my M.Sc programme.

Special note of thanks goes to the research group for their continuous moral support, accompany and friendly cooperation. Last but not the least, and most importantly, I wish to thank my parents and my wife, for being my driving force and standing by me through difficult time of this research. Without them, I would never have come so far in pursuing my dream.

Abstract

In this thesis, a noisy speech enhancement method based on noise compensation performed on short time magnitude as well phase spectra is presented unlike the conventional spectral subtraction method. Here, the noise estimate to be subtracted from the noisy speech spectrum is proposed to be determined exploiting the low frequency regions of noisy speech of current frame rather than depending only on the initial silence frames. We argue that this approach of noise estimation offers the capability of tracking the time variation of the non-stationary noise thus resulting in a noise compensated magnitude spectrum. By employing the noise estimates thus obtained, a procedure is formulated to compensate the distortion in the phase spectrum,which is kept unchanged in the typical speech enhancement methods. The noise compensated phase spectrum is then recombined with the noise compensated magnitude spectrum to produce a modified complex spectrum thus synthesizing an enhanced frame. Extensive simulations are carried out using NOIZEUS database in order evaluate the performance of the proposed method. It is shown in terms of objective measures, spectrogram analysis and informal subjective listening test that the proposed method consistently outperforms some of the state-of-the-art methods of speech enhancement from noisy speech corrupted by white or train or babble noise of very low levels of SNR.

Contents

List of Tables

List of Figures

.

Chapter 1

INTRODUCTION

The problem of improving performance of speech communication systems in noisy environments has been a challenging area for research for more than three decades. Now important applications of noise suppression and speech enhancement systems include improving the performance of 1) digital mobile radio telephony systems, which suffer both from background noise in the environment as well as from channel noise; 2) hands free telephone systems suffering from car noise etc.; 3) pay phones in a noisy environment (e.g. restaurants, factories, airports); 4) air-ground communication systems in which pilot's speech is corrupted by cockpit and engine noise; 5) ground-air communication where the cockpit/engine noise corrupts the received signal; 6) long distance communication over noisy radio channels; 7) teleconferencing systems where a noise source from one location maybe broadcasted to all other locations; and 8) hearing aids and cochlear implants in a noisy environment (e.g. classrooms, cafeteria etc. Efforts to achieve higher quality and intelligibility of noisy speech may effectively end up improving performance of other speech applications such as speech coding/compression and speech recognition etc.

The terms "speech enhancement" and "speech cleaning" properly refer respectively to the improvement of the quality or intelligibility of speech and the reversal of degradations that have corrupted it; in practice however, most authors use the two terms interchangeably. The principal degradations that we are concerned with are (a) additive acoustic noise, (b) acoustic reverberation, (c) convolutive channel effects resulting in an uneven or band limited response, (d) non-linear distortion such as arises from clipping, (e) additive broadband electronic noise , (f) electrical interference, (g) codec distortion , (h) distortion introduced by recording apparatus.

Speech enhancement methods attempt to improve the performance of communication systems when their input or output signals are corrupted by noise. The main objective of speech enhancement or noise reduction is to improve one or more perceptual aspects of speech, such as the speech quality or intelligibility. This is important in a variety of contexts, such as in environments with interfering background noise (e.g offices, streets and automobiles etc.) and in speech recognition systems. Over the year, researchers and engineers have developed a number of methods to address this problem. Yet, due to complexities of the speech signal, this area of research still poses a considerable challenge. It is usually difficult to reduce noise without distorting speech and thus, the performance of speech enhancement systems is limited by the tradeoff between speech distortion and noise reduction.

The overall goal of a speech enhancement technique is to reduce listener fatigue, to boost the overall speech quality, to increase intelligibility, and to improve the performance of the voice communication device [1], [2] . The goal varies according to specific application and each application has an aim to make a tradeoff between two or among several goals. In general, since the presence of noise seriously degrades the performance of the systems in such applications, the efficacy of the systems operating in a noisy environment is highly dependent on the speech enhancement techniques employed therein.Therefore,the aims of speech enhancement can be summarized as follows:

1. Improvements in the intelligibility of speech to human listeners.
2. Improvement in the quality of speech that make it more acceptable to human listeners.
3. Modifications to the speech that lead to improved performance of automatic speech or speaker recognition systems.
4. Modifications to the speech so that it may be encoded more effectively for storage or transmission.

1.1 Problem Definition

In speech communication system, during transmission and reception, signals are often corrupted by noise, which can cause severe problems for further processing

and user perception. The presence of background noise [3] in speech significantly reduces the intelligibility of speech. Degradation of speech severely affects the ability of person, whether impaired or normal hearing, to understand what the speaker is saying. Therefore an effective means of removing the noise from speech is invaluable for many speech processing applications by using noise reduction or speech enhancement algorithms that can suppress such background noise and improve the perceptual quality and intelligibility of speech. The speech signal can be acquired from single or multiple channel sensors. Single channel systems constitute one of the most difficult situations of speech enhancement, since no reference signal to the noise is available, and the clean speech cannot be preprocessed prior to being affected by the noise. It is already known that additive noise make speech degraded and non-stationarity of the noise process can further complicate the enhancement effort. One microphone input (single channel) could make speech enhancement difficult, as speech and noise are present in the same channel. Separation of the two signals would require relatively good knowledge of the speech and noise models or require that the interfering signal be present Exclusively in a different frequency band than that of the speech signal. Usually single channel systems make use of different statistics of speech and unwanted noise.The performance of these methods are usually limited in presence of non-stationary noise as most of the methods make an assumption that noise is stationary during speech intervals and also, the performance drastically degrades at lower signal to noise ratios. A costly solution to this problem is to use a dual channel microphone approach. These systems take advantage of the availability of multiple signal inputs to the system and make use of the noise reference in an adaptive noise cancelation device, the use of phase alignment to reject undesired noise components, or even the use of phase alignment and noise cancelation stages into a combined scheme. But these systems tend to be more complex and more costly.In general, the situation where the noise and speech are in the same channel (single channel systems) is the most common scenario and is one of the most difficult situations to deal with.The complexity and ease of implementation of any proposed scheme is another important criterion especially since the majority of the speech enhancement and noise reduction algorithms find applications in real-time portable systems like cellular phones, hearing aids, hands free

kits etc. Hence, between the two systems, single microphone systems are the most common real-time scenario e.g. mobile communication, hearing aids etc. as usually a second channel is not available in most of the applications. These systems are easy to build and comparatively less expensive than the multiple input systems. Therefore, this thesis is concerned with enhancement methods that use only a single microphone signal. Single channel speech enhancement methods can be generally divided into several categories based on their domains of operation, namely time domain, frequency domain and time-frequency domain. Time domain methods includes the subspace approach [4]- [8], frequency domain methods includes speech enhancement methods based on discrete cosine transform [9], the spectral subtraction [10], [11], minimum mean square error (MMSE) estimator [4], [12], [13], Wiener filtering [5], [14] and time frequency-domain methods involve the employment of the family of wavelet [15]- [20] and [8], which is a superior alternative to the analyses based on Short Time Fourier Transform (STFT). The main challenge in such denoising approaches based on the thresholding of the wavelet coefficients [21] of the noisy speech is adjusting the threshold value, so that it can prevent enhanced speech distortion as well as decrease residual noise. Then, by using the threshold, the designing of a thresholding scheme to minimize the effect of wavelet coefficients corresponding to the noise is another difficult task. In order to handle the practical situations of real life applications, a speech enhancement method, apart from providing simplicity in computation, is needed to be capable of producing optimal results with improved overall speech quality with minimized speech intelligibility loss under low levels of SNR.

Since the majority of the speech enhancement and noise reduction methods find applications in real-time portable systems like cellular phones, hearing aids, hands free kits etc., the complexity and ease of implementation of any proposed scheme is another important criterion. In order to attain these goals by reducing noise from the noisy speech, various speech enhancement methods namely, spectral subtraction (SS) [10], [11], MMSE estimation [22], [11] subspace [4]- [8] Wiener filtering [5], and Kalman filtering [6] have been reported in the literature.The spectral subtraction method has been one of the most well-known techniques for noise reduction. The spectral subtraction estimates the power spectrum of clean speech by explicitly sub-

tracting the noise power spectrum from the noisy speech power spectrum. Due to its mini-mal complexity and relative ease in implementation, it has enjoyed a great deal of attention over the past years. This approach generally produces a residual noise commonly called musical noise. Among all the methods mentioned above, although spectral subtraction suffers from an artifact known as musical noise, it has been widely used due to its noise suppression capability with simple computation. All the conventional speech enhancement methods, considers only the magnitude spectrum of the noisy speech while keeping the corresponding phase spectrum unchanged for synthesizing a cleaner speech. In [23], [24] the phase spectrum has been modified keeping the magnitude spectrum unchanged. Yet, the fact that at low levels of SNR, the changes in phase have indeed an effect on speech understandability, has not been studied in detail. Thus, in severe noisy conditions, development of a speech enhancement method incorporating noise compensations in both amplitude and phase spectra is still a challenging task.

Therefore, in this thesis, we intend to develop a speech enhancement method, where not only the short time magnitude is noise compensated but also the short time phase spectrum is altered to handle and reduce the noise causing unwanted distortion in the enhanced speech. We rely on the fact that noisy speech spectrum in low frequency regions is equivalent to the noise spectrum in that region. Therefore, unlike the conventional methods of SS, noise estimate is proposed to be determined exploiting the low frequency regions of noisy speech of the current frame rather than depending only on the initial silence frames and the determined noise spectrum estimate is updated in every silence period. Such an approach of noise estimation and resulting noise compensated magnitude spectrum offer the capability of tracking the time variation of the non-stationary noise. Unlike the majority of the conventional speech enhancement methods that keep the short time phase spectrum unchanged, we design a real value frequency dependent phase compensation function in order to modify the complex spectrum of the noisy speech. The angle of the modified complex spectrum of the noisy speech represents the noise-compensated phase spectrum. We propose the degree of phase spectrum compensation to be dependent on the magnitude of the noise spectrum estimate of the current frame thus allowing the phase spectrum also to follow the time variation of the non stationary noise.

Since the phase compensation function is anti-symmetric, it acts as the cause for changing the angular phase relationship. The noise compensated phase spectrum in conjunction with the noise compensated magnitude spectrum produces an enhanced complex spectrum that is found to contribute to cancel out low energy components more than the high energy components while performing enhanced speech synthesis thus reducing background noise drastically. Extensive simulations are carried out using NOIZEUS [25] database in order evaluate the performance of proposed method. It is shown in terms of objective measures and subjective evaluations that the proposed method consistently outperforms some of the state-of-the-art methods of speech enhancement from noisy speech corrupted by white or train or babble noise of very low levels of SNR. The proposed method is able to provide an improved overall speech quality with minimized speech intelligibility loss under low levels of SNR thus can be suitable to be employed in speech communication applications, such as mobile telephony, speech coding and recognition, and hearing aid devices involving practical noisy conditions.

1.2 Objective of the Thesis

The objectives of this thesis are:

1. To develop a noise compensation scheme for the magnitude spectrum based on the low frequency regions of noisy speech of the current frame.

2. To formulate a noise compensation scheme for the phase spectrum by employing the obtained noise estimate.

3. To investigate the performance of the proposed speech enhancement method in terms of objective and subjective measures using speech signals of a standard speech database in the presence of different noises in a wide range of SNRs from high to low.

The outcome of this thesis is a spectral enhancement method for noisy speech based on noise compensated magnitude and phase spectra thus synthesizing an enhanced speech with improved quality and minimal intelligibility under low levels of SNR.

1.3 Organization of this Thesis

This thesis is organized as follows; Chapter 1 gives the introduction of the overall thesis. Chapter 2 gives an overview of hearing and various aspects of hearing which are critical in developing noise reduction criteria based on human perception along with a review of various speech enhancement methods Chapter 3 discusses the proposed Modified Spectral Subtraction (MSS) approach with correction in phase.Results and simulations parts are elaborated in chapter 4. Finally, concluding remarks and suggestions for future works are provided in Chapter 5.

Chapter 2

LITERATURE REVIEW

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

In this chapter, an overview of hearing and various aspects of hearing which have been important in development of some of the recent perceptually based enhancement [26], [27] methods is presented here. Since real world is concerned with noises, different types of noises and their characteristics are described in detail in this chapter. Basically,this chapter presents an overview of different speech enhancement

methods, with greater emphasis on single channel subtractive type algorithms and provides a review of some of the major aspects and approaches in this category.

2.1 Overview of Hearing

The human auditory system has an unsurpassed capability to adapt to noise. There has been a great deal of research domain order to model this capability for purpose of enhancing speech. In the past few decades, good progress has been made in Understanding how the hearing mechanism works in processing a sound in general and Especially in context of noise.The human auditory system is based on a time-frequency analysis of sounds. The information received by the human ears can be described most conveniently as non-linear auditory responses to frequency selectivity and perceived loudness. The general properties of frequency selectivity are related to the concepts of critical band.Critical bands correspond to a physical measurement in the cochlea. Fletcher's band widening experiment laid the foundation for the critical-band concept by the assumption that incoming sounds are preprocessed by the peripheral auditory system through a bank of band pass filters. Each of these auditory filters acts like a frequency.Weighing function, corresponding closely to the frequency selectivity of the ear across the critical bands. The notion of critical band is related to the phenomenon of masking.Masking occurs because the auditory system is not able to differentiate two signals close in frequency or in time. It is manifested by a shift of auditory threshold in signal detectability. Loudness is another important attribute of auditory perception in terms of which sounds can be ranked on a scale extending from quiet to loud. These aspects of human auditory system such as critical band structure,masking, absolute threshold, excitation patterns etc. have been applied in speech coding,speech recognition and speech quality evaluation.The following section describes the above-mentioned aspects of human auditory perception, which addresses the basis of modeling perceptual properties andincorporating them into speech processing systems especially in the context of speech enhancement systems.

2.1.1 The Human Ear

The human auditory system consists of the ear, auditory nerve fibers, and a section of the brain. It converts sound waves into sensations perceived by the auditory cortex. The ear is the outer peripheral system, which converts acoustic energy (sound waves) into electrical impulses that are picked up by the auditory nerve. The ear itself is divided into three parts, the outer, middle and inner ear as shown in Fig. 2.1.

The Outer Ear

The outer ear consists of the pinna (the visible part of the ear), the meatus (ear canal), and terminates at the tympanic membrane, also known as the eardrum. The pinna is primarily responsible for collecting sound, and aids in sound localization. The meatus is a tube, which directs the sound to the tympanic membrane. A cavity with one end open and the other end closed by the tympanic membrane, the meatus acts as a quarter-wave resonator with a center frequency around 3000 Hz. This particular structure likely aids the perception of obstruents (sounds produced by obstructing the air flow in the vocal tract, such as $/s/$ and $/f/$), which have their energy content in this frequency region.

The Middle Ear

The middle ear is considered to begin at the tympanic membrane and contains the ossicles, a set of three small bones. The bones are named malleus (hammer), incus (anvil), and stapes (stirrup). Acting primarily as levers performing an impedance matching transformation (from the air outside the eardrum to the fluid in the cochlea),they also protect against very strong sounds. The acoustic reflex activates the middle ear muscles, to change the type of motion of the ossicles when low-frequency sounds with Sound Pressure Level (SPL) above 85-90 dB reach the eardrum. Attenuating pressure transmission by up to 20 dB, the acoustic reflex is also activated during voicing in the speaker's own vocal tract . Due to their mass, the ossicles act as a low-pass filter with a cutoff frequency around 1000 Hz.

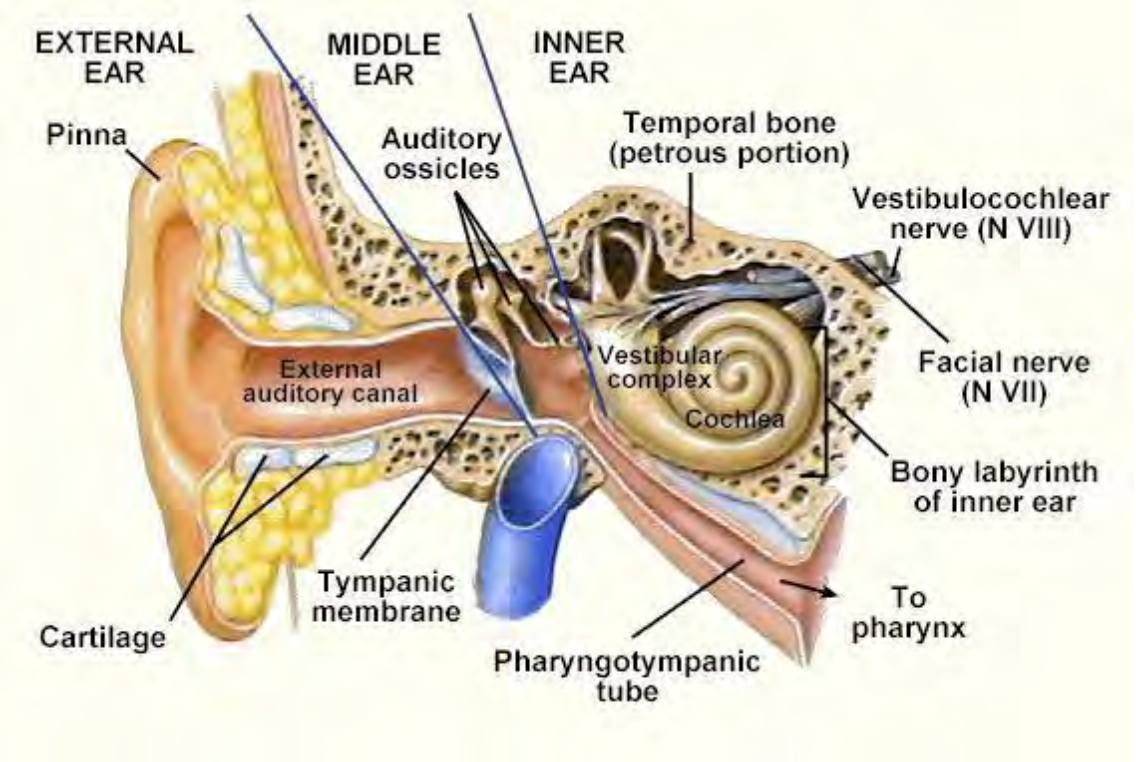

Fig. 2.1: Structure of Human Ear

The Inner Ear

The inner ear is a bony structure comprised of semicircular canals of the vestibula and the cochlea. The vestibula is the organ that helps balancing the body with no apparent role in the hearing process [18]. The cochlea is a cone-shaped spiral in which the auditory nerve terminates. It is the most complex part of the ear that transmits the mechanical oscillations to the basilar membrane through the fluid present in the inner ear, which produces very small displacements of the basilar membrane. The Corti, which is located on the basilar membrane and contains about 4×3500 hair cells (for a total length of 32mm) performs a frequency-place transformation of these mechanical oscillations into electrical pulses. Linear distance along the basilar membrane corresponds approximately to logarithmic increments of frequency.

2.1.2 The Critical Band

The critical band structure can be used to describe many aspects of the behavior of the auditory system. The critical band has a perceptual and a physical relationship

with the auditory system. A basic definition of the critical band is the "bandwidth at which subjective response changes rather abruptly". Another view of the critical band is that it represents a first approximation of the ear's ability to discriminate different frequencies. Experiments have shown that 25 critical bands exist over the frequency range of human hearing, which spans from 20 Hz to 20 kHz. It is evident that the critical bands have constant width of 100 Hz for center frequencies up to 500 Hz, and the bandwidths increase as the center frequency increases further. Since location on the basilar membrane has an approximately linear relationship to the frequency scale for low frequencies but a logarithmic relationship at higher frequencies, the linear frequency scale is inadequate for representing the auditory system. Critical band analysis is the basis for almost all the models based on auditory system. One critical band corresponds to a 1.5 mm step along the basilar membrane that contains 1200 primary auditory nerve fibers. Critical band analysis is the first stage of analysis performed by the inner ear. As mentioned earlier, this analysis is a frequency-domain transformation, which can be seen as a filterbank with bandpass filters. A critical filter bank, gives equal weight to portions of speech with the same perceptual importance.

2.2 Noise Characteristics

The nature of the noise is an important factor in deciding on a speech enhancement method. Therefore, a good model of noise is important for the performance of speech enhancement system and vice-versa it is important to analyze how well a speech enhancement algorithm/model works with different types of noise.Based on the nature and properties of the noise sources, noise can be classified in the following ways:

1. Background noise: additive noise, which is usually uncorrelated with the signal and present in various environment scenarios like cars, offices, city streets, fans,machines, climatic conditions, factory environment, cockpits, helicopters etc.From these types of noise, both noise (white noise filtered to model long-term average of room noise) is stationary, noises in streets and factories etc. have more dynamic characteristics. Factory and helicopter noise having strong periodic components and noise from fans, and car noise in a hands free envi-

ronment etc. are real noise and are examples of non-stationary noise having time varying characteristics.

2. Interfering speakers (speech like noise): additive noise, composed of single or multiple "competing" speakers. The multi-talker babble which also attributes to the phenomenon called "cocktail party effect"(many voices talking simultaneously, e.g. in a cafeteria, a noisy classroom) is noise, which has characteristics and frequency range very similar to the speech signal of interest.

3. Impulse noise: slamming of doors, noise present in archived gramophone recordings.

4. Non-additive noise: noise due to non-linearities of microphones, speakers and channel distortion (speech on transmission lines).

5. Non-additive noise due to speaker stress: e.g. Lombard effect i.e. the effect induced in presence of noise, wherein the speaker has a tendency to increase his vocal effort. This results in speech having different spectral properties as compared to clean speech, a detailed summary of the changes in speech characteristics due to this effect are given in Speech produced under situational and emotional stress also fall in this category.

6. Noise correlated with the signal: reverberations and echos.

7. Convolutive noise: corresponds to convolution in time domain. For instance, changes in speech signal due to changes in room acoustics or changes in microphones etc. These are usually harder to deal with, as compared to additive noise.

8. Multiplicative noise: signal distortion due to fading in cellular channels.

In general, it is more difficult to deal with non-stationary addative noise, where there is no priority knowledge available about the characteristics of noise. Since non-stationary noise is time varying, the conventional method of estimating the noise from initial intervals assuming no speech signal is not suitable for estimation [10]. Noise types, which are similar in temporal, frequency or spatial characteristics to speech, are also difficult to remove or attenuate. Multitalker babble, for instance,

retains some characteristics of speech and In general, it is more difficult to deal with non-stationary noise, where there is no priori knowledge available about the characteristics of noise. Since non-stationary noise is time varying, the conventional method of estimating the noise from initial intervals assuming no speech signal is not suitable for estimation. Noise types, which are similar in temporal, frequency or spatial characteristics to speech, are also difficult to remove attenuate. Multitalker babble, for instance, retains some characteristics of speech and poses a particularly difficult problem for an algorithm intended to isolate speech signal from the additive noise. Fig. 2.2 through Fig. 2.9 shows the effect of various type of noises on the clean speech signal. Fig. 2.2 shows the time domain representations of the clean speech which plots the amplitude of the clean speech signal with respect to time. Fig. 2.6 is the spectrograms of the clean speech, which provides a three dimensional representation of short speech utterances. From these figures we see that, the spectrograms display separate harmonics and they aid analysis of pith and vocal tract resonance. Due to limited range on spectrograms or to filtering of the speech,however, harmonics are often invisible . For better understanding of the effect of various types of noises on the clean speech signal both in the time and frequency domains, Fig. 2.2 through Fig. 2.5 are plotted for the time domain representations and Fig. 2.6 through 2.9 show the spectrogram representations of the speech signal.In this thesis we are concerned with additive stationary and nonstationary noise.

The majority of speech enhancement algorithms operate only the spectral magnitude. This stems from the generic principle “the ear is phase deaf”. This principle is only partially true.It is true that the phase can be changed drastically - as is the case in reverberant speech without greatly affecting understanding; thus the main effect of phase seems to rather qualitative. However, phase is not free either. We presume that random and fast changes in phase have indeed an effect on speech understandability [23]. However, this effect has not been studied in detail .In speech enhancement algorithms - especially those of the spectral subtraction nature the spectral magnitude is modified in order to remove as much of the noise as possible. Then the cleaned magnitude is recombined with the noisy phase to reconstruct a cleaned up speech signal. In this situation the phase and magnitude components are

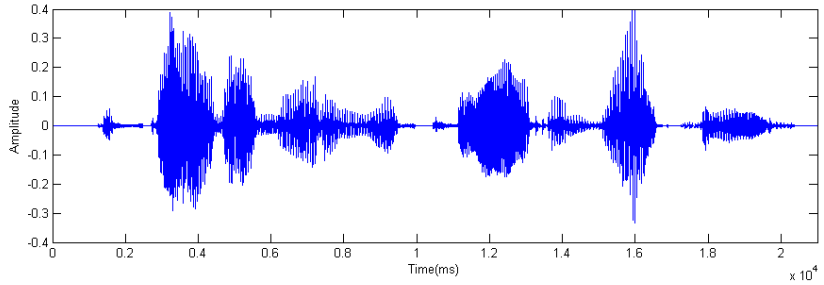

Fig. 2.2: Clean speech signal

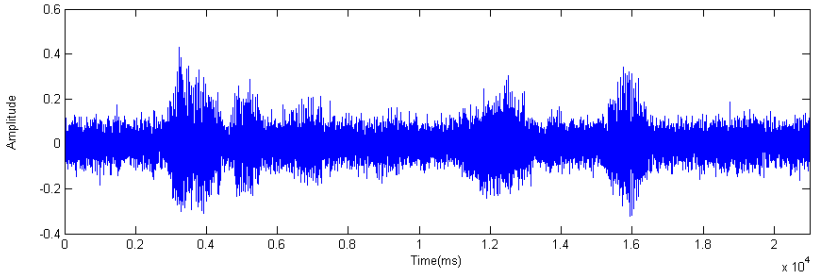

Fig. 2.3: 0 dB White noise corrupted speech signal

not necessarily consistent. Especially in speech resynthesis with an overlap-add [3] paradigm phase isn't free due to the usage of overlapping frames. Using an iterative algorithm first designed by Griffin and Jim phase can be adjusted to be as consistent as possible.

2.3 Classification of single channel speech enhancement Techniques

Speech Enhancement (SE) systems can be classified in a number of ways , based on the criteria used or application of the enhancement system. Typically, the speech en-

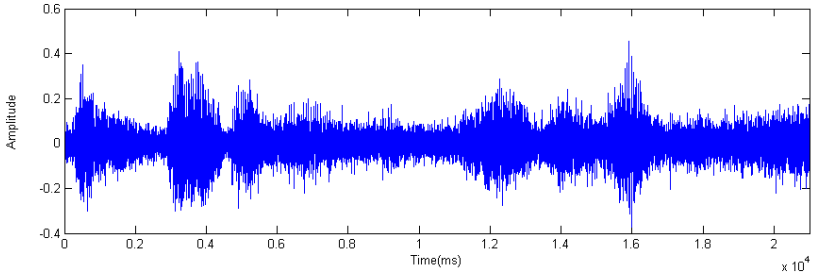

Fig. 2.4: 0 dB train noise corrupted speech signal

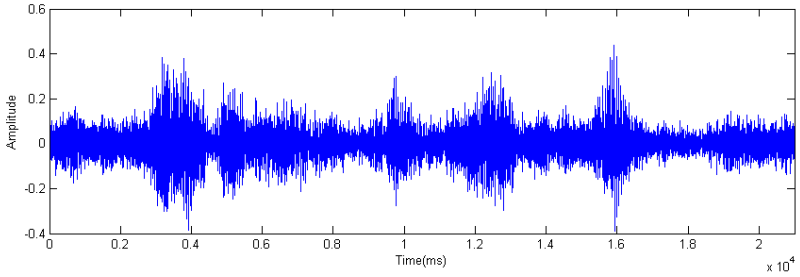

Fig. 2.5: 0 dB babble noise corrupted speech signal

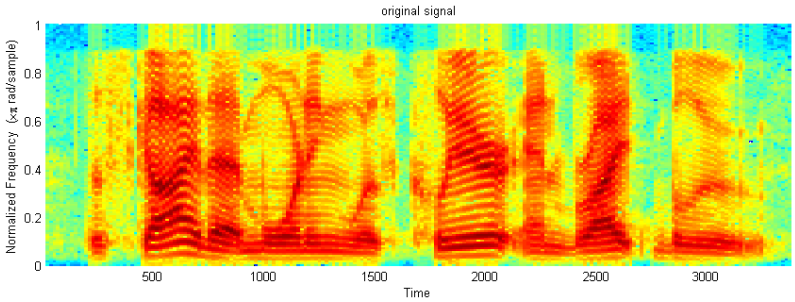

Fig. 2.6: Spectrogram of Clean speech signal

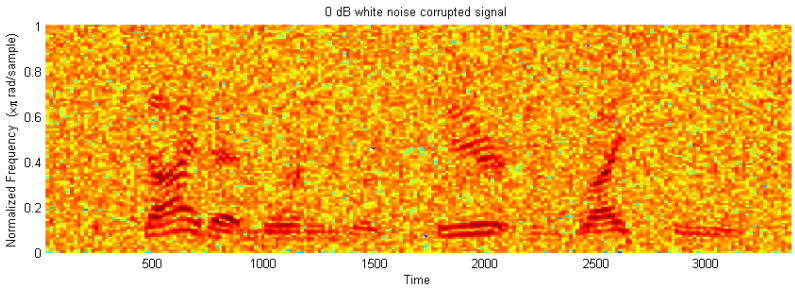

Fig. 2.7: Spectrogram of 0 dB white noise corrupted speech signal

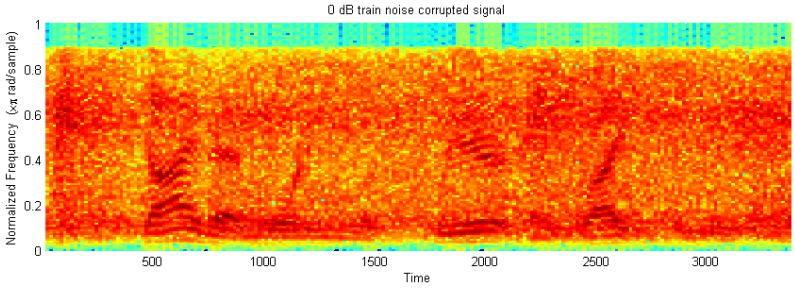

Fig. 2.8: Spectrogram of 0 dB train noise corrupted speech signal

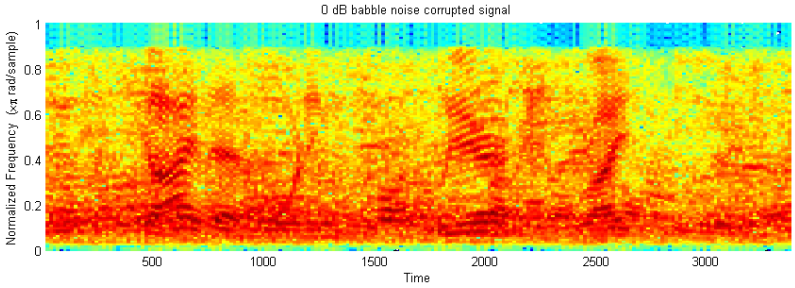

Fig. 2.9: Spectrogram of 0 dB babble noise corrupted speech signal

hancement literature broadly divides the various speech processing strategies under single and multichannel enhancement techniques. The performance of single channel systems is usually limited because they tend to improve the quality of the noisy signal at the expense of some intelligibility loss [28] [29]. Hence time, frequency, and time-frequency domain single channel speech enhancement methods, reported in the literature, have their own advantages and drawbacks. The discussion in this chapter will be limited to single channel enhancement techniques at different domains, as these are the most common types of enhancement systems found in many applications.

2.4 Overview of the time domain single channel SE methods

One of the popular time domain speech enhancement methods include the subspace method [4]- [8]. The idea behind the subspace approach is to project the noisy signal onto two subspaces: the signal-plus-noise subspace, or simply signal subspace (since the signal dominates this subspace), and the noise subspace. The noise subspace contains signals from the noise process only, hence an estimate of the clean signal can be made by removing the components of the signal in the noise subspace and retaining only the components of the signal in the signal subspace. Some well known methods for the decomposition of the space into two subspaces are the eigenvalue decompositions (EVD) [10], [6], [7]. The EVD based methods by Ephraim and Van Trees in [22] provided a noise shaping mechanism, but unfortunately they can only be applied to white noise and a separate noise whitening procedure needs to be utilized when colored-noise is present. The important succeeding extension to

colored noise by Mittal et al. in [18] is suboptimal methods since approximations were made to obtain the desired results. In [9], Jabloun and Champagne proposed a way to incorporate the auditory model into the subspace-based methods for speech enhancement. They first performed the eigen decomposition of the clean signal covariance matrix (by subtracting the noisy signal and noise covariance matrices) and then estimated the masking thresholds using the spectrum derived from an eigen- to frequency domain transformation. The masking thresholds were transformed back to the eigenvalue domain using a frequency-to-eigen domain transformation and then incorporated into the signal-subspace approach. The authors implicitly assumed that the subspace occupied by the speech signal is the same as that occupied by the signal obtained after the auditory model is taken into account. In the subspace method [9], a mechanism to obtain a tradeoff between speech distortion and residualnoise is proposed with the cost of a heavy computational load.

subspace approach

One particular class of speech enhancement techniques that has gained a lot of attention is signal subspace filtering. In this approach, a nonparametric linear estimate of the unknown clean-speech signal is obtained based on a decomposition of the observed noisy signal into mutually orthogonal signal and noise subspaces. This decomposition is possible under the assumption of a low-rank linear model for speech and an uncorrelated additive (white) noise interference. Under these conditions, the energy of less correlated noise spreads over the whole observation space while the energy of the correlated speech components is concentrated in a subspace thereof. Also, the signal subspace can be recovered consistently from the noisy data. Generally speaking, noise reduction is obtained by nulling the noise subspace and by removing the noise contribution in the signal subspace. It is assumed that the original signal exhibits some well-defined Properties or obeys a certain model. Signal enhancement is then obtained by mapping the observed signal onto the space of signals that possess the same structure as the clean signal. This theory forms the basis for all subspace-based noise reduction algorithms. A first and indispensable step towards noise reduction is obtained by nulling the noise subspace (least squares (LS) estimator) [3].Of particular interest is the minimum variance (MV) estimation, which gives the best linear estimate of the clean data, given the rank p of the clean

signal and the variance of the white noise [4], [5]. Later on, a subspace-based speech enhancement with noise shaping was proposed in [6]. Based on the observation that signal distortion and residual noise cannot be minimized simultaneously, two new linear estimators are designed-time domain constrained (TDC) and spectral domain constrained (SDC)-that keep the level of the residual noise below a chosen threshold while minimizing signal distortion. Parameters of the algorithm control the trade-off between residual noise and signal distortion. In subspace base speech enhancement with true perceptual noise shaping, the residual noise is shaped according to an estimate of the clean signal masking threshold, as discussed in more recent papers [7]-[9]. Although basic subspace-based speech enhancement is developed for dealing with white noise distortions, it can easily be extended to remove general coloured noise provided that the noise covariance matrix is known (or can be estimated)

2.5 Overview of the Frequency domain single channel SE methods

2.5.1 Spectral subtraction approach

Spectral subtraction [10], [13], [28], [30], [31], [32] is a well-known noise reduction method based on the short term spectral amplitude (STSA) estimation technique [10]- [13]. The basic power spectral subtraction technique, as proposed by Boll [10], is popular due to its simple underlying concept and its effectiveness in enhancing speech degraded by additive noise. The basic principle of the spectral subtraction method [10], [13], [28], [30], [31], [32] is to subtract the magnitude spectrum of noise from that of the noisy speech. The noise is assumed to be uncorrelated and additive to the speech signal. An estimate of the noise signal is measured during silence or non-speech activity in the signal. A general representation of the technique is given in Fig. 2.10. The enhanced signal has largely reduced noise levels compared to the original noise corrupted signal resulting in a better SNR and improved speech quality. However, although spectral subtraction method is simple and provides a tradeoff between speech distortion and residual noise to some extent, it suffers from an artifact known as “musical noise” having an unnatural structure that is perceptually annoying, composed of tones at random frequencies and has an increased variance. It is obvious that the effectiveness of the noise removal process is

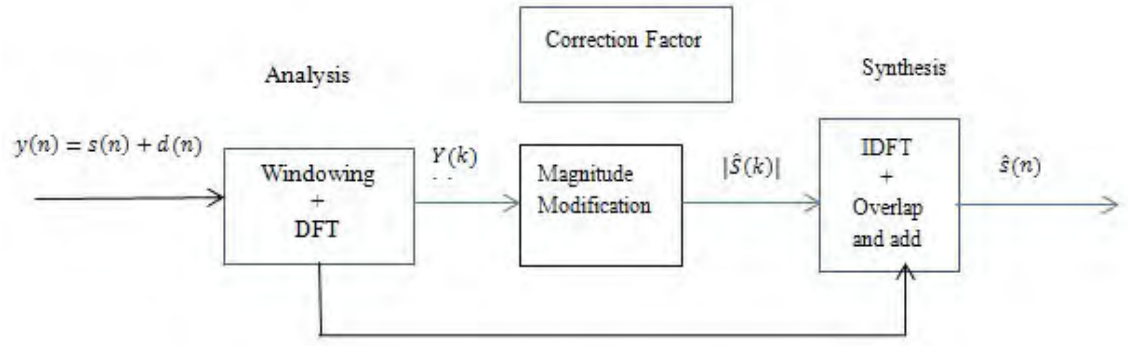

Fig. 2.10: Diagrammatic representation of short time spectral magnitude enhancement system

dependent on obtaining an accurate spectral estimate of the noise signal. The better the noise estimate, the lesser the residual noise content in the modified spectrum. However, since the noise spectrum cannot be directly obtained, we are forced to use an averaged estimate of the noise. Hence there are some significant variations between the estimated noise spectrum and the actual noise content present in the instantaneous speech spectrum. The subtraction of these quantities results in the presence of isolated residual noise levels of large variance. These residual spectral content manifest themselves in the reconstructed time signal as varying tonal sounds resulting in a musical disturbance of an unnatural quality. This musical noise can be even more disturbing and annoying to the listener than the distortions due to the original noise content. This and other drawbacks of the method neutralize the improvement in speech quality achieved due to the reduction in noise levels and can be more annoying than the original noise itself. In Boll's method [10] of spectral subtraction (SS), the noise spectrum is estimated from the non-speech frames and subtracted from the noisy speech spectrum in the current frame. If the noise is stationary, then the noise estimation becomes accurate and the resulting spectrum on transforming it in the time domain produces a cleaner speech. However, in practice, since most of the noises are usually nonstationary [33], the Boll's SS method results in a degraded performance in terms of speech quality and intelligibility. The method in [34] proposed by Paliwal, where noise is estimated based on the high-order Yule-Walker equations without finding the non-speech frames, can track the non-stationary noise but needs larger computations. The minimum statistics based

SS approach [23], [4] can handle non-stationary noise with lesser computations, but the outcome of the approach relies on the spectra estimated from the past frames. Recently, in Yamauchi's method [18] the noise spectrum is estimated in the current frame using high-frequency regions, where human speech information is absent. It is to be noted that this method needs a very high sampling rate, which is unrealistic in the context of speech processing systems. To this end, a SS method has been proposed in [7], where the spectrum of non-stationary noise is estimated without using the higher sampling rate. All the speech enhancement methods mentioned above, considers only the magnitude spectrum of the noisy speech while keeping the corresponding phase spectrum unchanged for synthesizing a cleaner speech.

Correction of phase in spectral subtraction

Several methods employ the analysis-modification-synthesis (AMS) framework [35]. Let us consider an additive noise model,

$$x(n) = s(n) + d(n), \tag{2.1}$$

where $x(n)$, $s(n)$ and $d(n)$ denote discrete-time signals of noisy speech, clean speech and noise, respectively. Since speech can be assumed to be quasi-stationary, it is analyzed frame-wise in the AMS framework through short-time Fourier analysis. The discrete short-time Fourier transform (DSTFT) of the corrupted speech signal $x(n)$ is given by

$$X(n,k) = \sum_{-\infty}^{\infty} x(m)w(n-m)e^{-\frac{j2\pi mk}{N}}, \tag{2.2}$$

where k denotes the kth discrete-frequency of N uniformly spaced frequencies and w(n) is an analysis window function. In speech processing, the Hamming window with 20-40 ms duration is typically employed. Using DSTFT analysis we can, subject to constraints described in [7], represent Eqn. 2.2 as

$$X(n,k) = S(n,k) + D(n,k), \tag{2.3}$$

where$X(n,k)$, $S(n,k)$, and $D(n,k)$ are the DSTFTs of noisy speech, clean speech, and noise, respectively. Each of these can be expressed in terms of the DSTFT

magnitude spectrum and the DSTFT phase spectrum. For instance, the DSTFT of the noisy speech signal can be written in polar form as

$$X(n,k) = |X(n,k)|e^{j\angle X(n,k)}, \tag{2.4}$$

where $|X(n,k)|$ denotes the magnitude spectrum and $\angle X(n,k)$ denotes the phase spectrum.

Traditional AMS-based speech enhancement methods modify only the magnitude spectrum while keeping the noisy phase spectrum unchanged for synthesis. This stems from the generic principle that "the ear is phase deaf". In [23], the phase spectrum has been modified keeping the magnitude spectrum unchanged. In the phase domain approach the noisy phase spectrum is modified leaving the noisy magnitude spectrum unchanged. Noise suppression is achieved by altering the DSTFT phase spectrum in such a way as to induce large synthesis cancellation among noise components during the inverse DSTFT operation.

2.5.2 Minimum mean square error (MMSE) estimator approach

In the MMSE estimator [4], [12], [13], the frequency spectrum of the noisy speech is modified to reduce the noise from noisy speech in the frequency domain. A relatively large variance of spectral coefficients is the problem of such an estimator. While adapting filter gains of the MMSE estimator, spectral outliers may emerge, that is especially difficult to avoid under noisy conditions. Unlike magnitude averaging where averaging is performed irrespective of whether the frame contains speech or noise, the proposed MMSE estimator performs non-linear smoothing only when the SNR is low, i.e. when the frame predominantly contains noise. The residual noise present due to this technique has been observed to be colorless. The method reduces the distortions in the speech parts due to averaging.

2.5.3 Wiener filtering approach

The Wiener filter is a popular adaptive technique that has been used in many enhancement methods [5], [36], [37]. The basic principle of the Wiener filter is to estimate an optimal filter from the noisy input speech by minimizing the Mean

Square Error (MSE) between the desired signal $s(k)$ and the estimated signal $\widehat{s}(k)$. The Wiener filter can be given in the frequency domain by:

$$H(w) = \frac{P_s(w)}{P_s(w) + P_n(w)}, \tag{2.5}$$

where $P_s(w)$ is the power spectral density (PSD) of the speech and $P_n(w)$ is the PSD of the noise spectrum calculated during periods of non-speech activity. From 2.1 it is obvious that a priori knowledge of the speech and noise power spectra is necessary. The speech power spectrum is estimated using the estimated speech model parameters. One of the major problems of wiener filter based methods is the requirement of obtaining clean speech statistics necessary for their implementation. Both the MMSE and the Wiener estimators have a moderate computation load, but they offer no mechanism to control tradeoff between speech distortion and residual noise.

2.6 Conclusion

In this chapter, brief literature surveys of the recent state-of-the-art speech enhancement [1] methods are presented. All the methods have their pros and cons. In order to handle the practical situations of real life applications, a speech enhancement method, apart from providing simplicity in computation, is needed to be capable of producing optimal results with improved overall speech quality with minimized speech intelligibility loss under low levels of SNR. Despite many relatively successful attempts to implement speech enhancement system in severe noisy conditions, development of a single approach of speech enhancement that offers the know-how of determining an appropriate threshold value as well as designing an effective thresholding scheme is still an open problem.

Chapter 3

A Spectral Domain Speech Enhancement Method Based on Noise Compensations in Both Magnitude and Phase Spectra

3.1 Introduction

For speech enhancement, it is well known that spectral subtraction has been widely used due to its noise suppression capability with simple computation. Most of the variations of spectral subtraction assume that noise remains stationary over time and considers only the modified magnitude spectrum while keeping the phase spectrum of noisy speech unchanged for synthesizing an enhanced speech. Although there are complex issues that involve the effect the phase spectrum on human audition, recently, it is suggested that phase spectrum can be useful for improved speech processing tasks [38]- [40].

In this chapter, a new approach to speech enhancement is developed, where not only the short time magnitude is noise compensated but also the short time phase spectrum is altered to handle the noise causing unwanted distortion in the enhanced speech. Based on the fact that noisy speech spectrum in low frequency region is equivalent to the noisy spectrum in that region, a noise estimation approach is introduced with the conditional spectral subtraction method in order to track the time variation of non-stationary noise. Unlike the conventional speech enhancement methods that help the short time phase spectrum unchanged, we proposed to incorporate the estimate noise spectrum in a procedure of noise compensation in the phase spectrum. the new complex spectrum obtained by exploiting the modified

magnitude and phase spectra is found effective is producing enhanced speech with improved quality with minimal distortion as compared to some of the existing speech enhancement methods [41].

3.2 PROBLEM FORMULATION

In the presence of additive noise $s_v[n]$, a clean speech signal $s_x[n]$ gets contaminated and produces noisy speech $s_y[n]$. The proposed method is based on the AMS framework where, speech is analyzed frame wise since it can be assumed to be quasi-stationary. The noisy speech is segmented into overlapping frames by using a sliding window . A windowed noisy speech frame is expressed in the time domain as

$$y[n] = x[n] + v[n], \tag{3.1}$$

where, $x[n]$ and $v[n]$ represent the windowed version of the clean speech $s_x[n]$ and that of the noise $s_y[n]$, respectively. In a transform domain, such as frequency domain, eqn. 3.1 can be expressed as

$$Y[k] = X[k] + V[k], \tag{3.2}$$

where, $Y[k]$, $X[k]$ and $V[k]$ are the frequency domain representations of frames of noisy speech , clean speech and noise in that order. In this paper, short-time Fourier transform (STFT) is employed to process and modify the noisy signal. The N-point STFT , $Y[k]$ of $y[n]$ can be computed as

$$Y[k] = \sum_{n=0}^{N-1} y[n] e^{-\frac{j2\pi nk}{N}} \quad if, 0 \leq k \leq N-1., \tag{3.3}$$

According to the AMS framework, first, the noisy speech frame $y[n]$ in processed in the transformed domain, then, modifications are carried out in the transformed domain and finally, the inverse transform of the operation followed by the overlap-add [3] synthesis is performed to reconstruct an enhanced speech frame. An overview of the proposed speech enhancement method is shown by a block diagram in Fig.3.1.In the AMS framework, we propose to employ the idea of SS based speech enhancement method due to its several attractive features, where, the short-time Fourier transform $Y[k]$ of $y[n]$ as expressed in eqn. 3.3 can be written in polar form as

$$Y[k] = |y[k]| e^{j\angle Y[k]}. \tag{3.4}$$

In eqn. 3.4, $Y[k]$ denotes the short-time magnitude spectrum and $\angle Y[k]$ denotes the short-time phase spectrum. In speech analysis, it is commonly believed that human auditory system is phase-deaf i.e., it ignores phase spectrum and considers only magnitude spectrum. That is why in the conventional spectral subtraction (SS) based speech enhancement methods, for synthesizing a clean speech, operations are performed only on the short-time magnitude spectrum and an unaltered short-time phase spectrum is maintained or vice versa. Recently, it has been shown that the phase spectrum is also useful in speech analysis [42], [24]. Therefore, in the intended SS based noise reduction scheme, we are inspired to modify not only the magnitude spectrum but also alter the phase spectrum of the noisy speech with a view to handle and reduce noise of different characteristics.

3.3 PROPOSED METHOD

3.3.1 Noise Compensated Magnitude Spectrum

In this section, the magnitude spectrum of the noisy speech is modified by exploiting the low-frequency regions of the noisy speech. Unlike conventional methods, the noise spectrum estimate is updated in every silence period and the low frequency regions of magnitude spectrum is taken into special consideration in order to compensate for the noise spectrum errors that may be induced in the spectral subtraction procedure specially when the additive noise is non-stationary changing its amplitude drastically with time. By using eqn. 3.3 and eqn. 3.4, the instantaneous power spectrum of $y[n]$ can be estimated as

$$|Y[k]|^2 \approx |X[k]|^2 + |V[k]|^2, \tag{3.5}$$

Since in a noisy environment, we do not have access to $x[n]$, we would like to obtain an estimate of $|X[k]|^2$ from $|Y[k]|^2$ For this purpose, a Fast Fourier transform (FFT) based power spectral subtraction scheme is derived from eqn. 3.5 as

$$|\widehat{X}[k]|^2 = \begin{cases} H[k] & if, H[k] > 0 \\ \beta_s |\widehat{V}[k]|^2 & otherwise, \end{cases} \tag{3.6}$$

where,

$$H[k] = |Y[k]|^2 - \alpha |\widehat{V}[k]|^2. \tag{3.7}$$

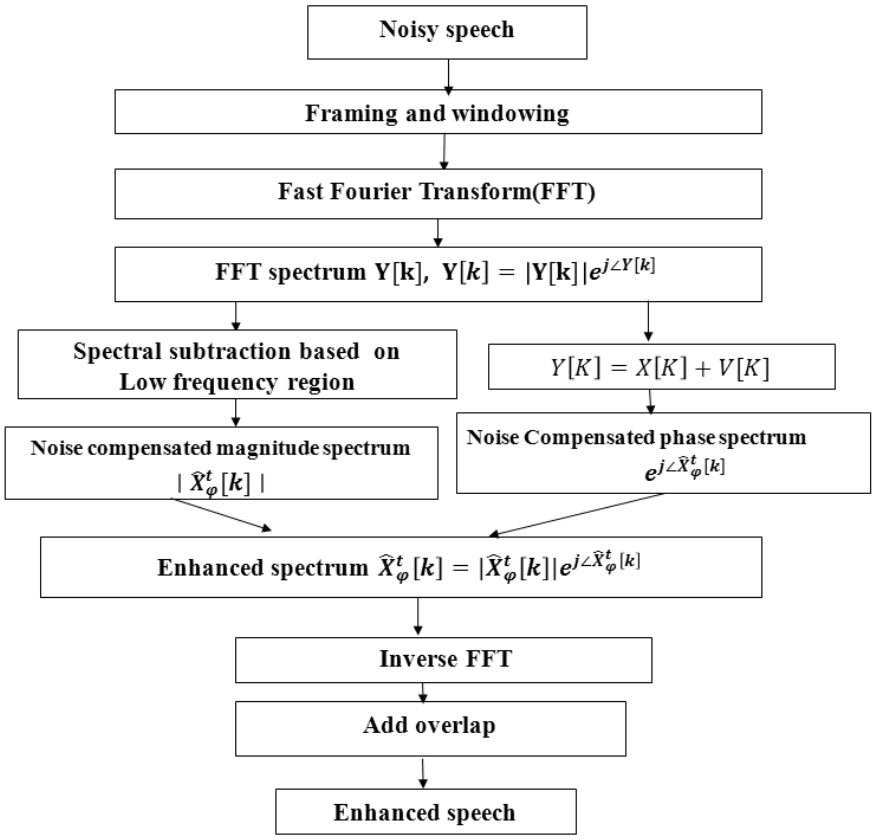

Fig. 3.1: Block diagram of proposed speech enhancement method

In eqn. 3.6, $|\widehat{X}[k]|^2$ represents at estimate of the short time FFT power spectra of $x[n]$ and β_s refers to the spectral flow parameter introduced to prevent the negative value of $|\widehat{X}[k]|^2$. In eqn. 3.7 , α symbolizes the over-subtraction factor used to prevent the overestimation of the noise power spectram. Conventionally an estimate $|\widehat{V}[k]|^2$ of the noise power spectrum $|V[k]|^2$is obtained from the beginning silence frames . In the proposed SS based noise reduction scheme, the noise power spectrum estimated from the beginning silence frames is updated during each silence period as follows

$$|\widehat{V}^t[k]|^2 = \begin{cases} |Y^t[k]|^2 & \text{for t=1} \\ v_n|V^{t-1}[k]|^2 + (1-v_n)|Y^t[k]|^2 & \text{otherwise} \end{cases}, \tag{3.8}$$

where,

$$|\widehat{V}^t[k]|^2 = |\widehat{V}^{t_I}[k]|^2,, \tag{3.9}$$

where t is the frame index, v_n is the forgetting factor, $|\widehat{V}^t[k]|^2$ and $|Y^t[k]|^2$, respectively, represent the estimated power spectrum and the power spectrum of the noisy speech at the t-th frame. After a silence period, for any frame t in the speech region, we rely on the use of a preliminary noise power spectrum estimated from eqn. 3.8 as eqn. 3.9 where t_I refers to the index of the immediate last silence frame before the beginning of a speech frame. Considering that this estimate of the noise power spectrum is updated only during a silence period while it may change drastically with time, it is insufficient to use a constant value of the over subtraction factor α to compensate for the errors induced in the noise power spectrum to be subtracted from the noisy speech power spectrum at each frame. In order to track the time variation of the noise , α should be adjusted at each frame t after a silence period. According to the spectral chateristics of human speech, the low frequency band typically from 0 to 50 Hz contains no speech information. Thus, for noisy speech, the low frequency band, say $\Delta = [0, 50]Hz$ contains only noise. In view of this fact, n order to change the value of α for the t-th frame after a silence period we propose to use the ratio between the powers of $|Y^t[k]|$ and $|\widehat{V}^{t_I}[k]|^2$ in low frequency band delta as

$$\alpha_t = \frac{\sum_{k\varepsilon\Delta}|Y^t[k]|^2}{\sum_{k\varepsilon\Delta}|\widehat{V}^{t_I}[k]|^2} \quad where \quad \Delta = [0, 50]Hz, \tag{3.10}$$

where $|\widehat{V}^{t_I}[k]|^2$ represents the estimated noise power in the immediately last silence

frame t_I before the beginning of speech frame. In the low frequency band Δ of the t th frame the variation of the noise speech power spectrum is equivalent to the noise power spectrum of that frame. Thus , use of α_t defined in eqn. 3.10 clearly serves as a relative weighing factor with respect to the estimated preliminary noise power spectrum $|\widehat{V}^t[k]|^2 = |\widehat{V}^{t_I}[k]|^2$, leading to a reasonable tracking for the time variation of the noise if nonstationary. Thus the noise compensated magnitude spectrum can be written as

$$|\widehat{X}^t[k]|^2 = \begin{cases} |Y^t[k]|^2 - \alpha_t|\widehat{V}^t[k]|^2 & if(|Y^t[k]|^2 - \alpha_t|\widehat{V}^t[k]|^2) > 0 \\ \beta_s|\widehat{V}^t[k]|^2 & otherwise. \end{cases} \tag{3.11}$$

3.3.2 Noise Compensated Phase Spectrum

In this section, by exploiting the noise estimate of the current frame as obtained in the previous section, the complex spectrum of the noisy speech is modified in such a way that the low energy component cancel out more than the high energy components. The modified complex spectrum thus obtained is used to obtain an altered phase spectrum that in conjunction with the noise compensated magnitude spectrum contribute to noise suppression while performing clean speech synthesis operation via FFT.

The noisy speech signal in the current frame $y^t[n]$ in the analysis stage is a real valued signal and therefore, its FFT is conjugate symmetric, ie.

$$Y^t[k] = \{Y^t[N-K]\}^*. \tag{3.12}$$

The conjugate can be obtained as a result of applying FFT on $y^t[n]$. The conjugate arise naturally from the symmetry of the magnitude spectrum and anti-symmetry of the phase spectrum . During IFFT operation as needed for clean speech synthesis, the conjugate are summed together to produce larger real valued signal. If the conjugate are modified, the degree to which they sum together can be influenced and this can be contributed constructively or destructively to the reconstruction of the clean time domain signal. In our approach, an ideal of the degree to which the conjugates reinforce or cancel during IFFT operation is by altering their angular relationship. Moreover, we propose the degree of phase spectrum compensation to

be dependent on the magnitude of the noise spectrum estimate of the current frame thus facilitating the handling of time and frequency varying no stationary noise conditions. For this purpose, we formulate a phase spectrum compensation function as given by

$$\phi[k] = \eta\Lambda[k]|\widehat{D}^t[k]| \quad where, \widehat{D}^t[k] = \alpha_t|\widehat{V}^t[k]|, \tag{3.13}$$

where,in eqn. 3.14 η is a real valued empirically determined constant, $|\widehat{D}^t[k]|$ is an estimate of the short time magnitude spectrum of noise in the current frame , and $\Lambda[k]$ presents a weighting function expressed as

$$\Lambda[k] = \begin{cases} 1 & ,if \quad 0 < \frac{k}{N} < \frac{1}{2} \\ -1 & ,if \quad \frac{1}{2} < \frac{k}{N} < 1 \\ 0 & otherwise \end{cases} \tag{3.14}$$

Here, zero weighting is assigned to the values of k corresponding to the non-conjugate vectors of FFT, such as $k = 0$ and value at $k = \frac{N}{2}$ if N even. Since the estimate of noise magnitude spectrum $\widehat{D}^t[k]|$ is symmetric, introduction of the weighting function $\Lambda[k]$ defined by 3.14 produces an anti-symmetric compensation function $\phi[k]$ that acts as the cause for changing the angular phase relationship in order to achieve noise cancelation during synthesis. A more in depth vector based explanation for two cases of single conjugate pair and their corresponding modifications are presented in Fig.3.2, where both the time frequency indexes are omitted for convenience and clarity. For the representation in Fig.3.2(a), the magnitude of the conjugates, ie. $\overrightarrow{Y}$ and $\overrightarrow{Y^*}$ are considered larger than that of the $\phi[k]$.Column one of Fig.3.2(a) shows the conjugate vectors $\overrightarrow{Y}$ and $\overrightarrow{Y^*}$ as well as their summation vector $\overrightarrow{Y + Y^*}$, in column two the real part of the $\overrightarrow{Y}$ and $\overrightarrow{Y^*}$ are shown to be offset by $|\phi|$ and -$|\phi|$, respectively.Thus alters the angles of the vectors $\overrightarrow{Y}$ and $\overrightarrow{Y^*}$ while keeping their magnitude unchanged thus Producing vectors $\overrightarrow{S_\phi}$ and $\overrightarrow{S'_\phi}$, respectively. It is seen from the column three that the vector $\overrightarrow{S_\phi + S'_\phi}$, is produced as a result of adding the modified vectors $\overrightarrow{S_\phi}$ and $\overrightarrow{S'_\phi}$. column four demonstrates the real part of the addition vector $\overrightarrow{S_\phi + S'_\phi}$, while its imaginary part is discarded with a view to avoid getting complex time domain frames after IFFT operation. Comparing column one and four of Fig.3.2(a), it is clear that a limited change of original signal occurs if $|\overrightarrow{Y}|$ and $|\overrightarrow{Y^*}|$ are greater than $|\phi|$. In Fig.3 2(b), similar illustration is shown by considering $|\overrightarrow{Y}|$ and $|\overrightarrow{Y^*}|$ is smaller than $|\phi|$ and found that significant change of

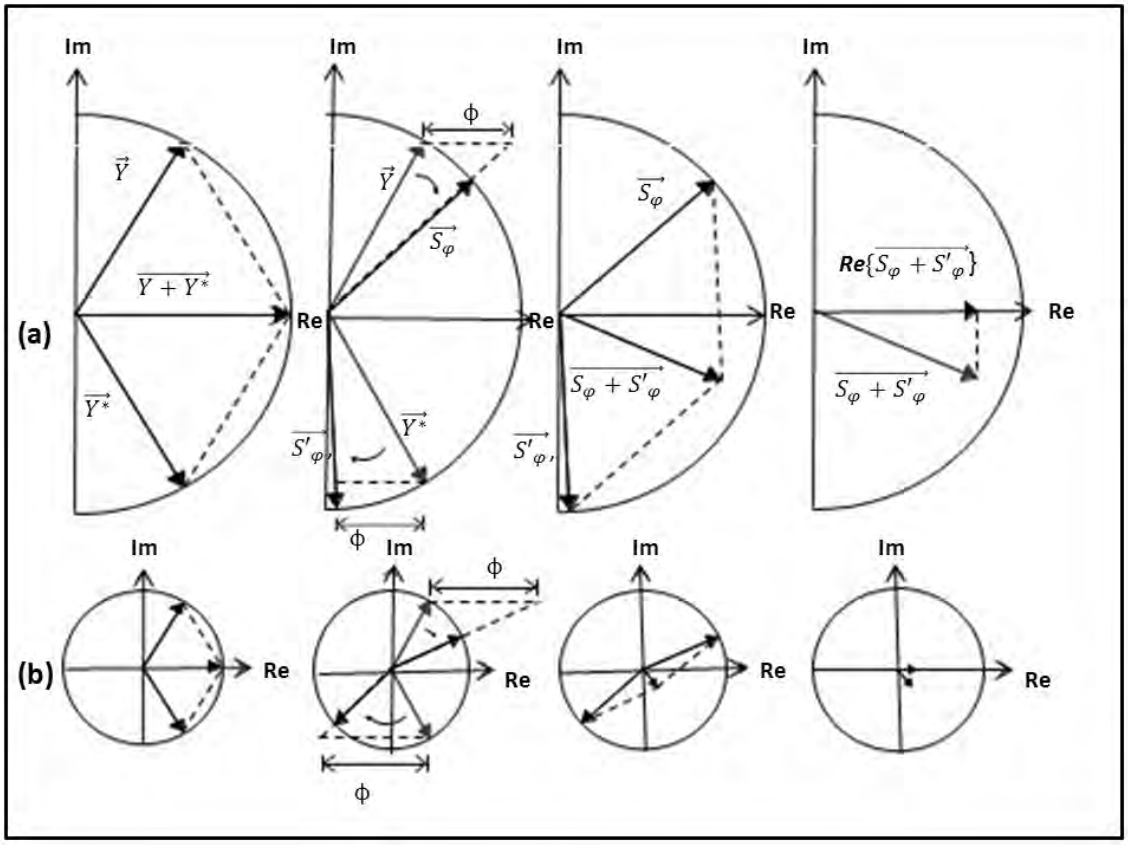

Fig. 3.2: Vector diagram of modification of conjugate symmetry.

the original signal occurs. Since $\phi[k]$ is anti symmetric, the angle of the conjugate pair in each case of Fig.3.2 are pushed in opposite directions, one towards 0 radian and other towards π radian. the Further they are pushed apart, the more out of phase they become. This justifies that, FFT spectrum of noisy speech with larger magnitude undergoes less attenuation and that with smaller magnitude subject to more attenuation based on the fact that noise frequency components are assumed to have lower magnitude than signal , when FFT spectrum of noisy speech has larger magnitude components. Using this assumption, which is basis for many noise cancellation and noise estimation algorithms, we compute the real value frequency dependent compensating function $\phi[k]$ and utilize it to offset the complex spectrum of the noisy speech as

$$\widehat{X}_\phi^t[k] = Y^t[k] + \phi^t[k]. \tag{3.15}$$

The strength of the compensation is dependent on the magnitude of both the FFT involving $Y^t[k]$ vectors and the $\phi[k]$ function. Finally , the noise-compensated phase

spectrum is obtained from $\widehat{X}_{\phi}^{t}[k]$ as

$$\angle\widehat{X}_{\phi}^{t}[k] = ARG[\widehat{X}_{\phi}^{t}[k], \tag{3.16}$$

where ARG is complex angle function. Although the compensated phase spectrum may not possesses the properties of phase spectrum of the clean speech it is cable of tracking phase compensating required due to noise present in each frame by incorporating the noise estimate $\widehat{D}^{t}[k]$ of corresponding frame while constructing the compensating function $\Lambda[k]$ used for computing the compensated phase spectrum.

3.3.3 Resynthesis of Enhanced Signal

In the synthesis stage , the compensated phase spectrum recombined with the compensated magnitude spectrum in order to produce an enhanced complex spectrum.

$$\widehat{X}^{t}[k] = |\widehat{X}^{t}[k]e^{j\angle\widehat{X}_{\phi}^{t}[k]}. \tag{3.17}$$

The enhanced speech frame is synthesized by performing the inverse FFT on the resulting $\widehat{X}^{t}[k]$,

$$\widehat{x}[n] = IFFT\{\widehat{X}^{t}[k]\},, \tag{3.18}$$

where $\widehat{x}[n]$ represents the the enhanced speech frame. The final enhanced speech signal is synthesized by using the standard overlap and add method [3].

3.4 Conclusion

A spectral enhancement method for noisy speech based on noise compensation both in magnitude and phase spectra is presented. Here a new spectral subtraction method is developed to obtain a noise compensated amplitude spectrum obtained by using a noise spectrum estimated by exploiting the low frequency regions of noisy speech of the frame under consideration. Such estimates of noise spectrum are incorporated to a proposed new procedure of phase-compensation required to handle phase distortion caused by additive noise. The modified complex spectrum obtained by recombining the noise compensated magnitude and phase spectra is found to cancel out of low energy components thus reducing background noise drastically.

Chapter 4

SIMULATION RESULT AND PERFORMANCE EVALUATION

In this chapter, a number of simulations are carried out to evaluate the performance of the proposed method.

4.1 Simulation Conditions and other Details

A noisy speech corpus (NOIZEUS) was developed to facilitate comparison of speech enhancement methods among research groups [17]. Thirty sentences from the IEEE sentence database [43] were recorded in a sound-proof booth using Tucker Davis Technologies (TDT) recording equipment. The sentences were produced by three male and three female speakers. The IEEE database (720 sentences) was used as it contains phonetically-balanced sentences with relatively low word-context predictability. The thirty sentences were selected from the IEEE database so as to include all phonemes in the American English language. The sentences recorded for NOIZEUS database [17] were originally sampled at 25 kHz and downsampled to 8 kHz. The noise was taken from the AURORA [44] database and includes suburban train noise, babble, car, exhibition hall, restaurant, street, airport and train-station noise. The noise signals taken from the AURORA database are artificially added to the clean speech signals in order to develop the noisy speech corpus.

We employ real speech sentences from the NOIZEUS noisy speech corpus for the evaluation of the performance of the proposed and other comparison methods. In our evaluation, in order to imitate a noisy environment, three different types of stationary and non-stationary noises, such as white, train and multi-talker babble noise are used. We have adopted train and babble noise corrupted speech from

the NOIZEUS noisy speech corpus. We have also used white noise available in the NOISEX92 [35] database to corrupt the NOIZEUS clean speech signals at different signal to noise ratio(SNR) levels.We consider noisy speech signals ranging from 15 db to 0 db for our simulations. Fig. 4.1 to Fig 4.8 show the spectrograms of a original clean speech signal uttered by a male and a female speakers and that of the same signals corrupted by white, train and babble noise of different SNRs (0 to 15). It is seen from these figures that harmonics representation in the original clean speech spectrogram is distorted while the different noises are added.

In order to obtain the overlapping analysis frames, hamming windowing operation is performed, where each frame is of 32 ms(256 samples), with a frame shift of 4 ms. 1024-point FFT operation is employed in this thesis in order to keep the noise estimation error minimal.

Figs. 4.9 through 4.11 demonstrate the setting of α_t over the frames of a sentence and the resulting time domain waveforms obtained by using the above settings in the proposed enhancement method in the presence of white, train and babble noise, respectively. It is seen from Figs. 4.9(a),(c); 4.10(a),(c) and 4.11(a),(c) that the α_t used in the proposed method is able to track the amplitude of different types of noise, particularly, the non-stationary case thus leading to a significant reduction of noise. This is validated by comparing noisy and enhanced time domain waveforms plotted in Figs. 4.9(b),(d); 4.10(b),(d) and 4.11(b),(d).

The value of η in 3.14 is determined empirically. Fig. 4.12 shows the empirical mappings of η as function of input speech SNR in dB for white ,train and babble noises. such empirical mapping are performed using all thirty utterance of the NOIZEUS database. The empirical are determined in a way such that the performance matrices used in our evaluation are maximized.

4.1.1 Comparison Metrics

Standard objective metrics, namely segmental SNR improvement in dB and perceptual evaluation of speech quality(PESQ) are employed for the evaluation of proposed method. The SNR of speech signal can be calculated as

$$SNR = 10log_{10} \sum \frac{x[n]^2}{v[n]^2}, \tag{4.1}$$

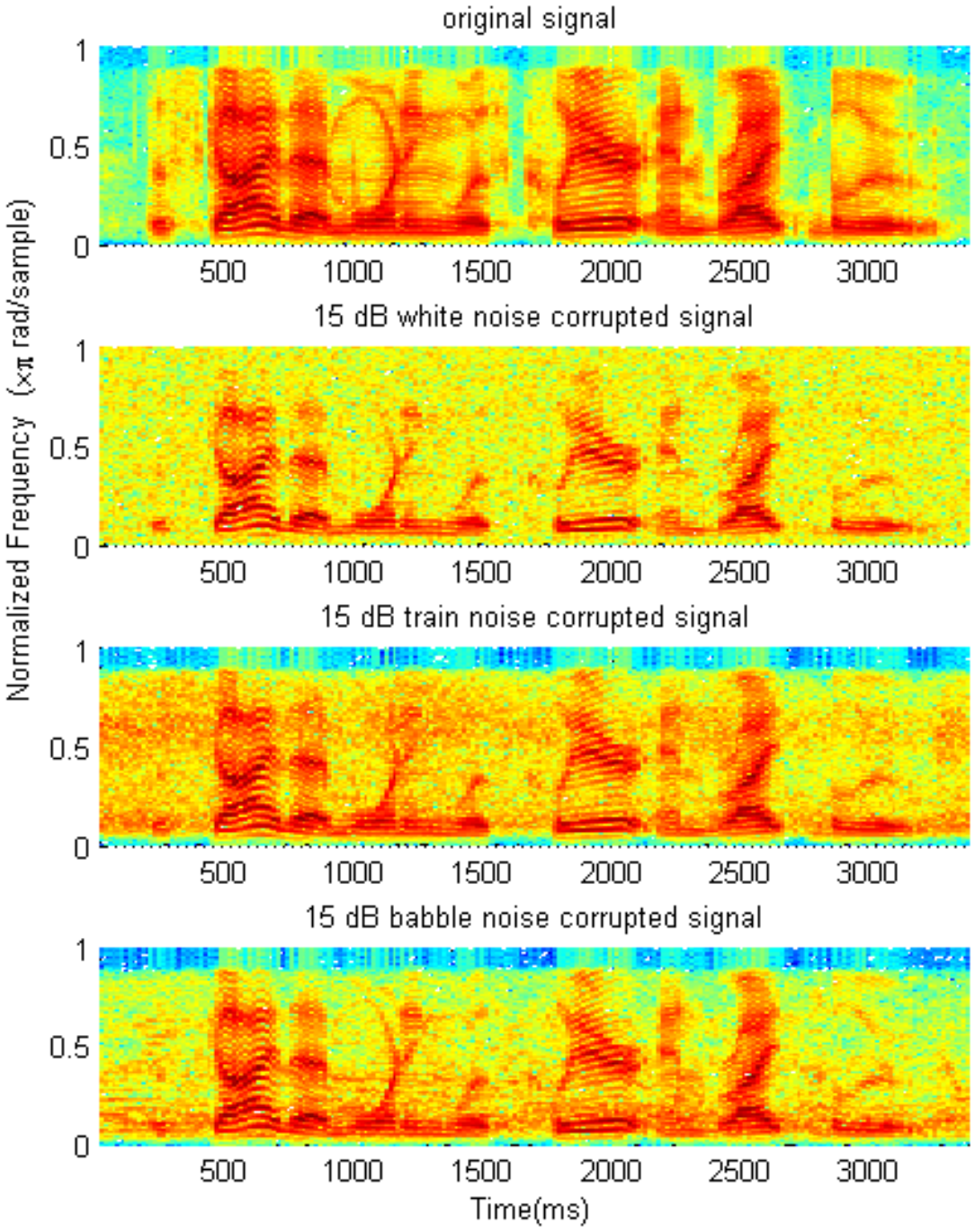

Fig. 4.1: Spectrogram of an original clean speech uttered by a male speaker and that of signals corrupted by white, train and babble noises at an SNR of 15 dB.

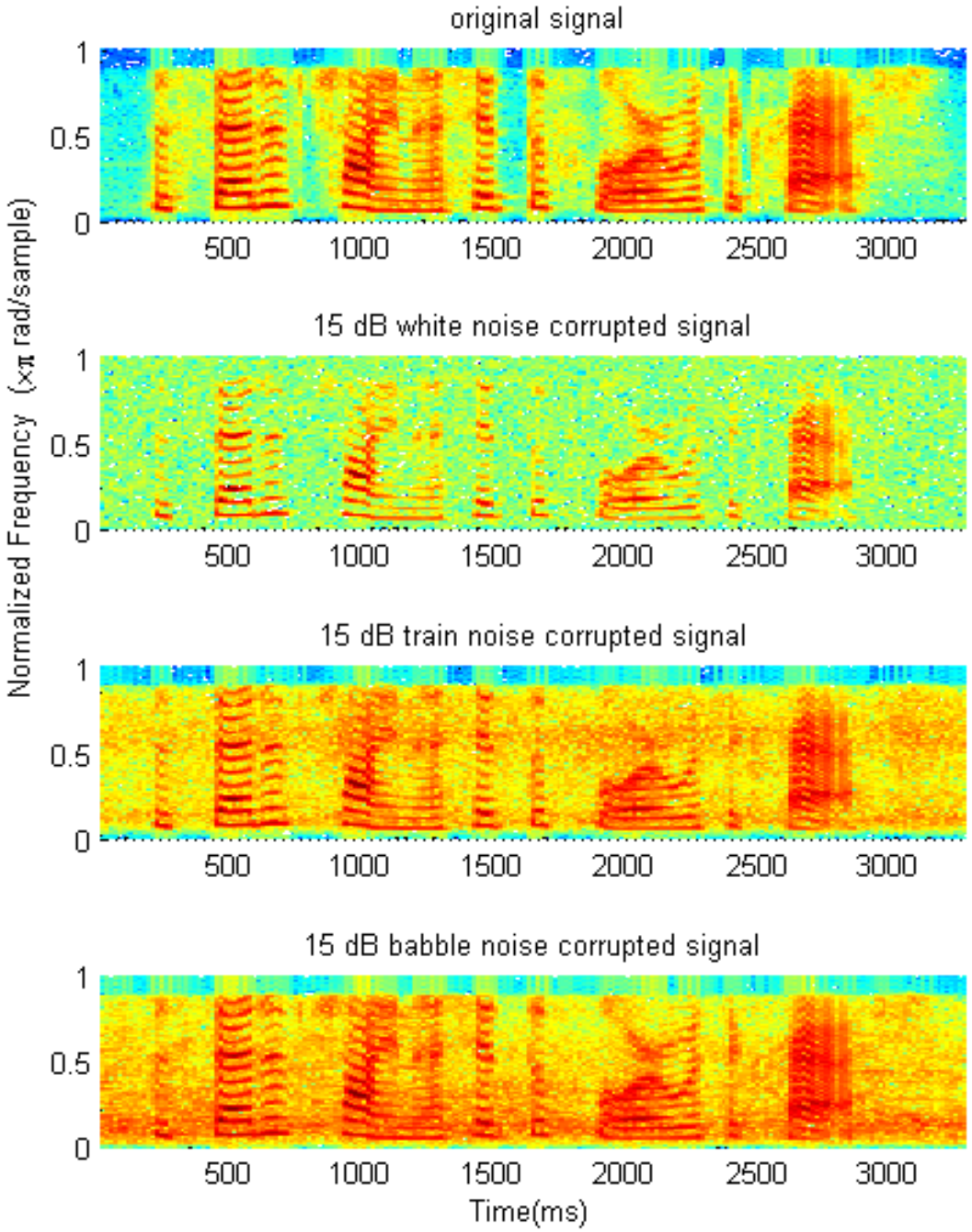

Fig. 4.2: Spectrogram of an original clean speech uttered by a female speaker and that of signals corrupted by white, train and babble noises at an SNR of 15 dB.

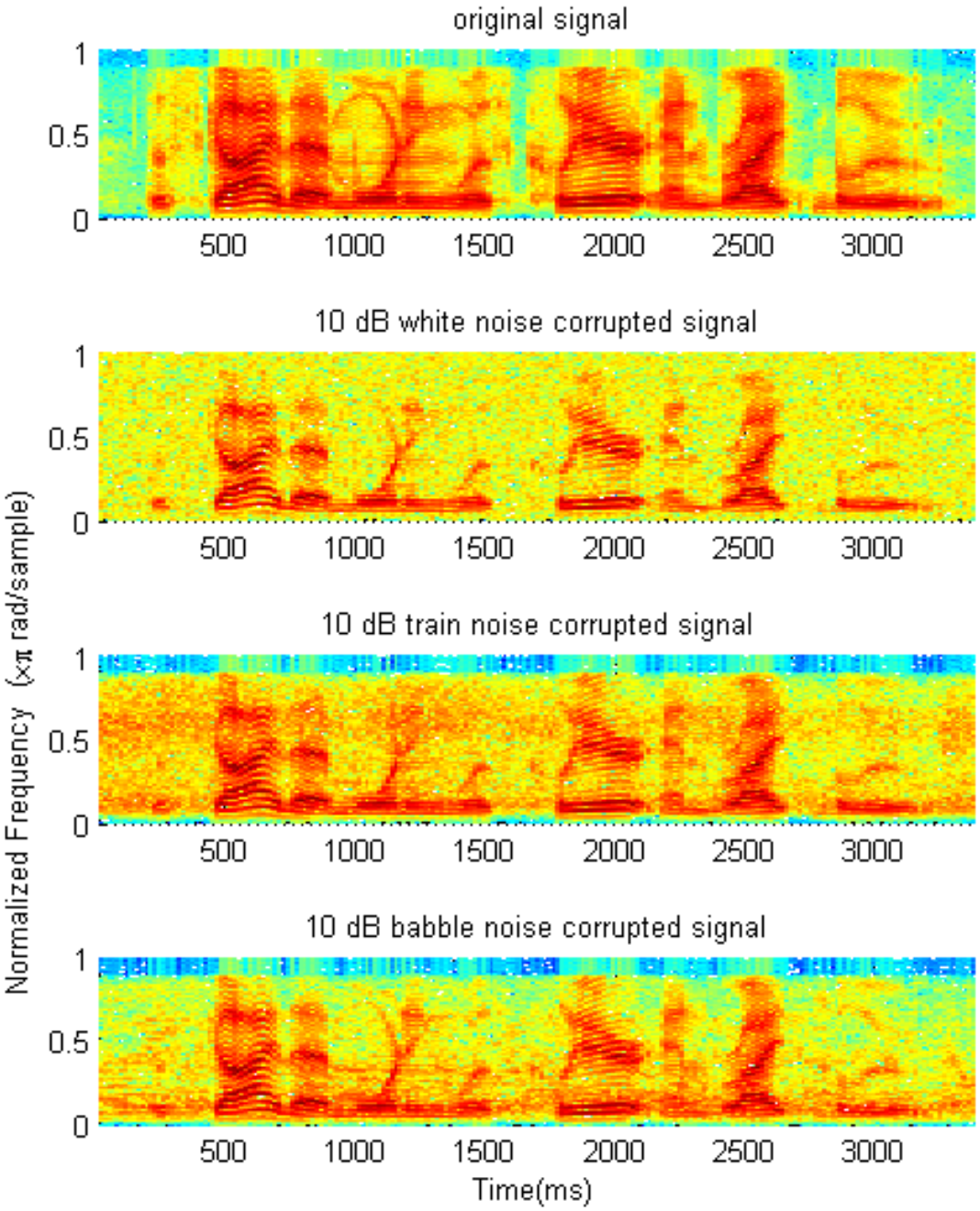

Fig. 4.3: Spectrogram of an original clean speech uttered by a male speaker and that of signals corrupted by white, train and babble noises at an SNR of 10 dB.

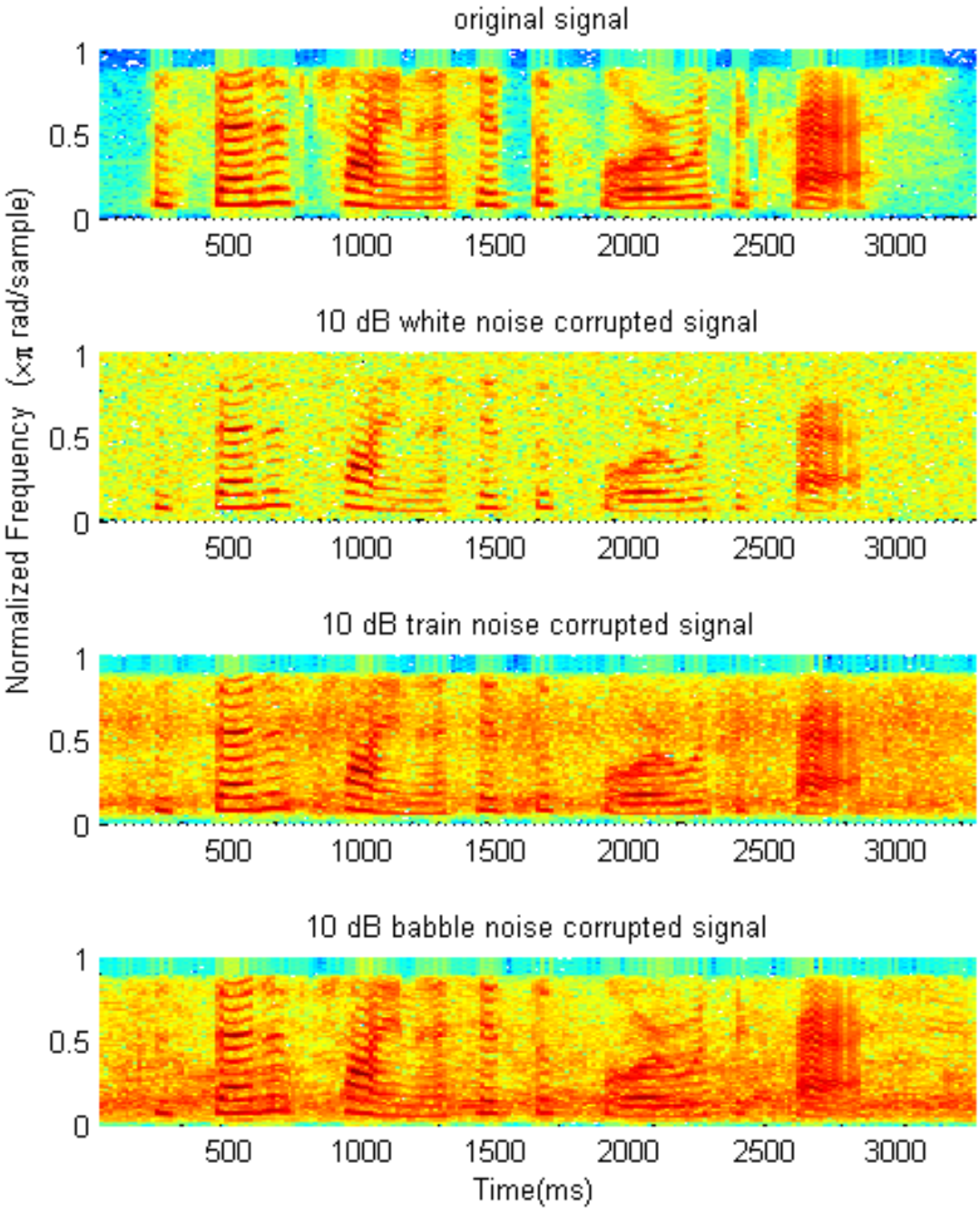

Fig. 4.4: Spectrogram of an original clean speech uttered by a female speaker and that of signals corrupted by white, train and babble noises at an SNR of 10 dB.

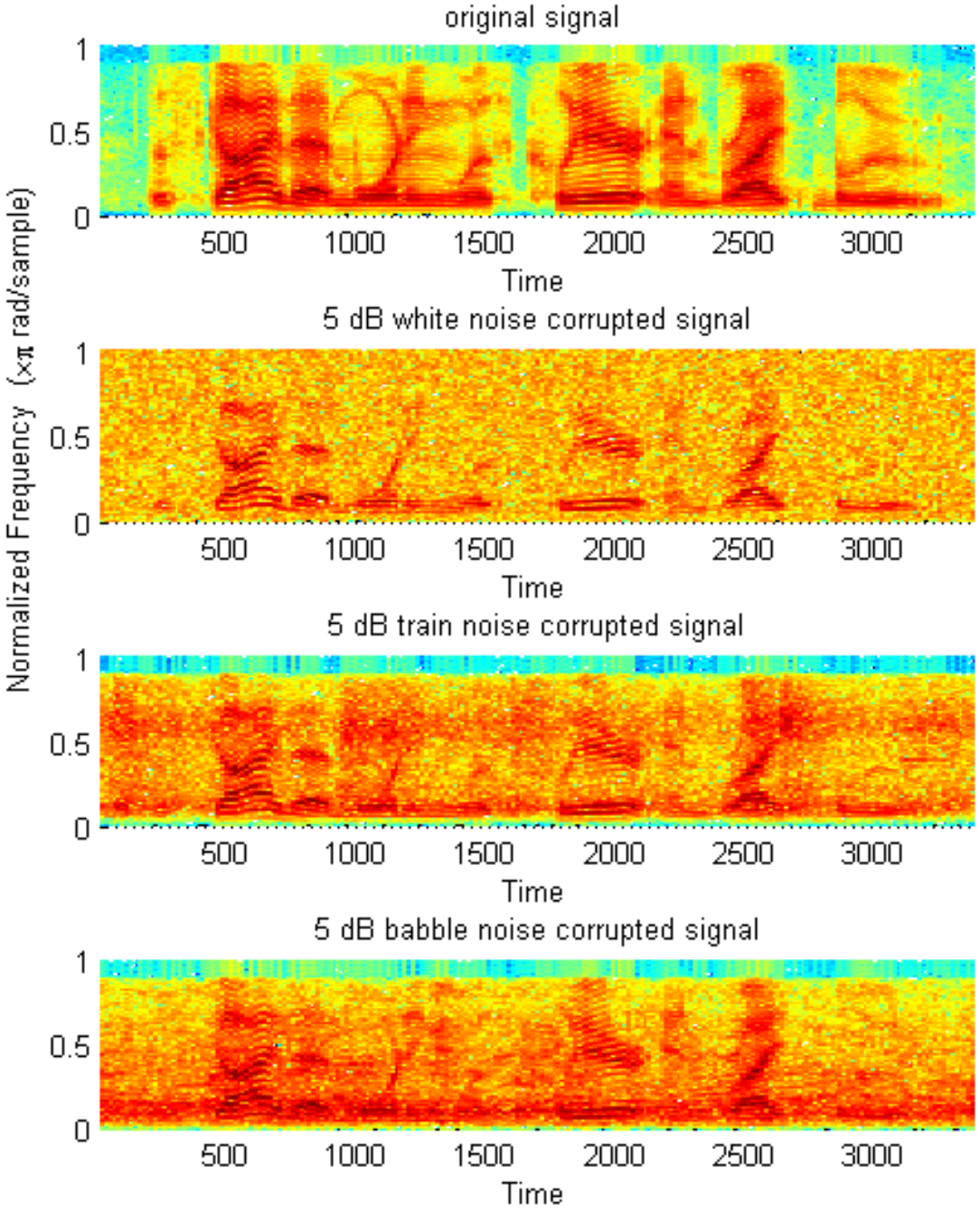

Fig. 4.5: Spectrogram of an original clean speech uttered by a male speaker and that of signals corrupted by white, train and babble noises at an SNR of 5 dB.

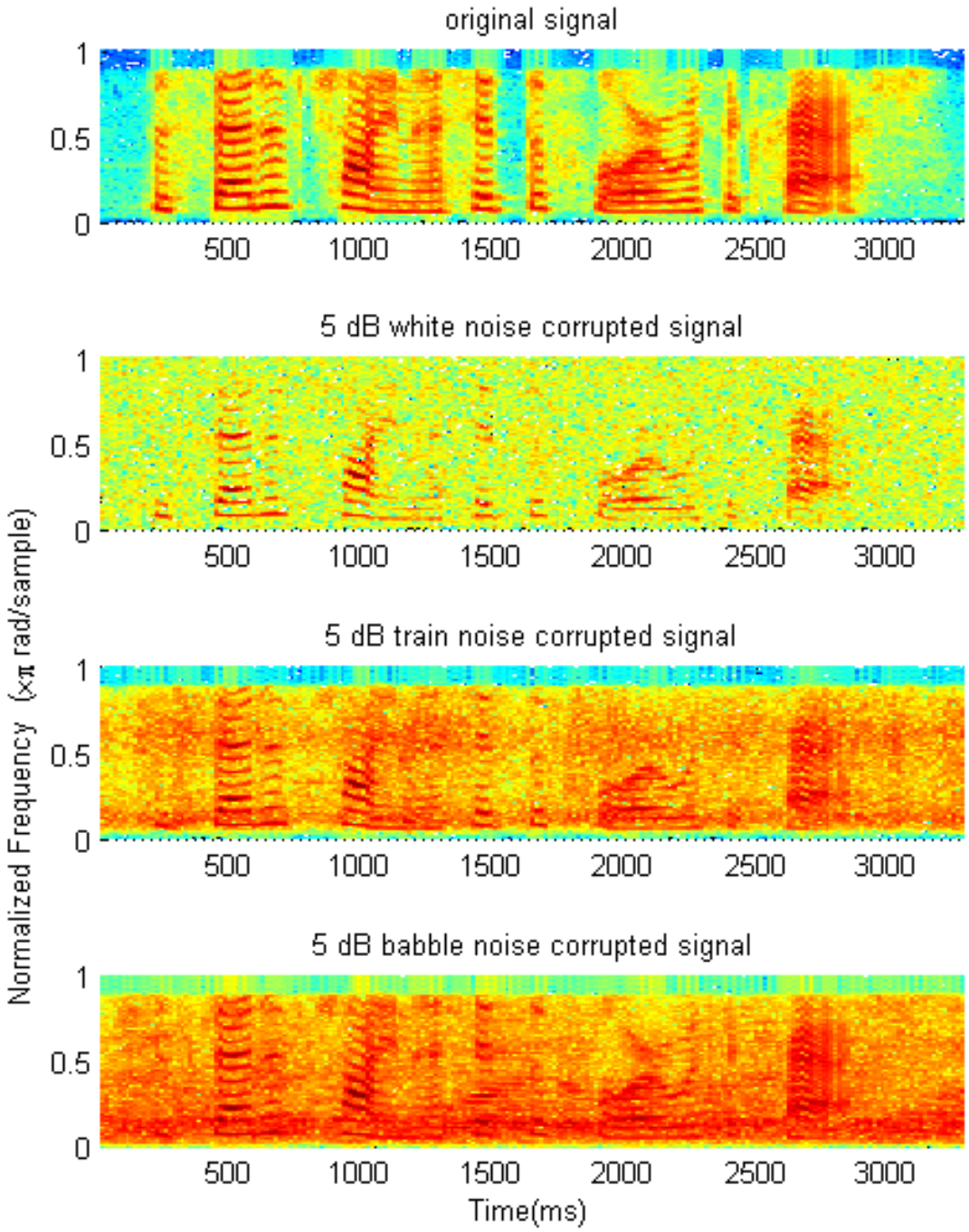

Fig. 4.6: Spectrogram of an original clean speech uttered by a female speaker and that of signals corrupted by white, train and babble noises at an SNR of 5 dB.

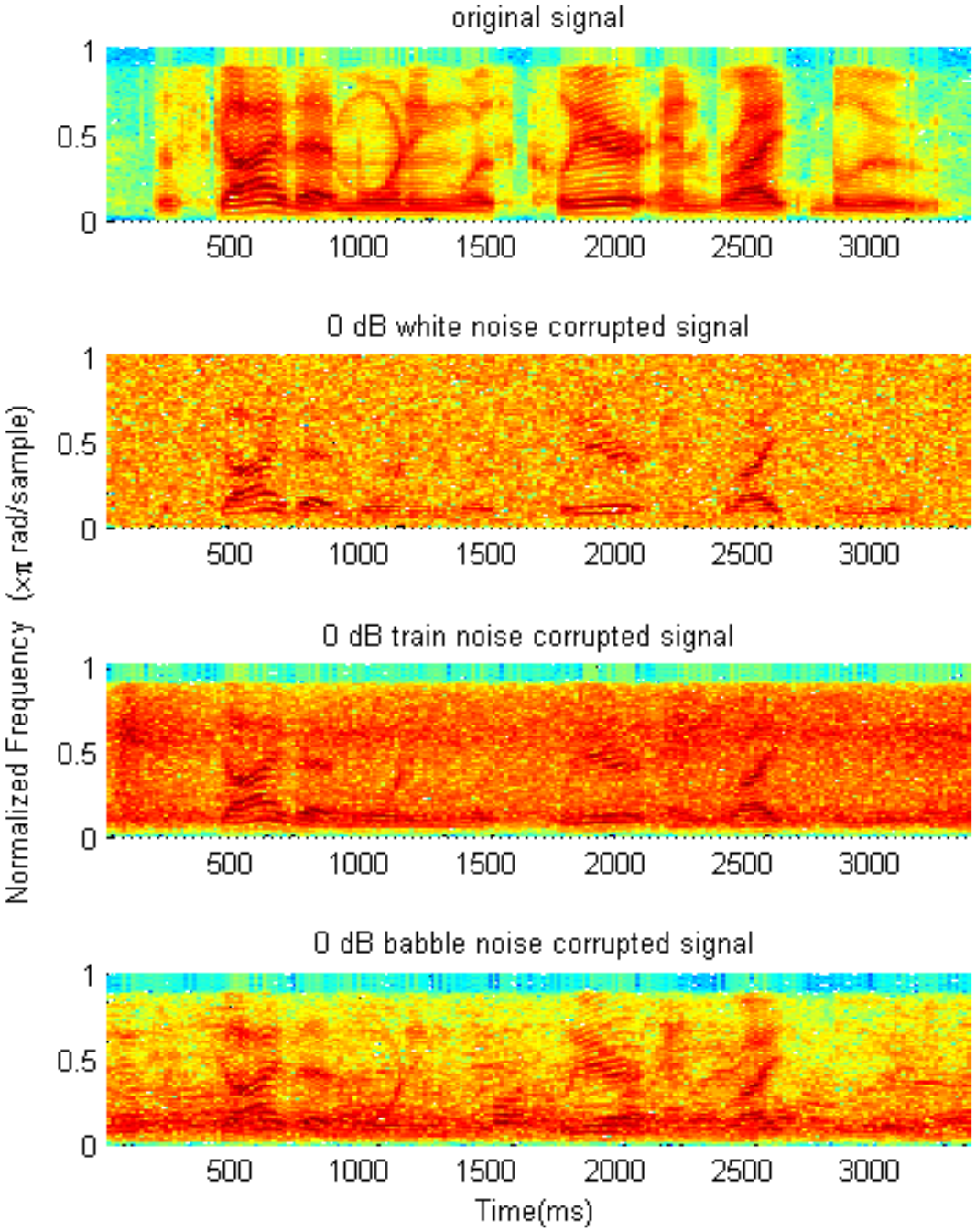

Fig. 4.7: Spectrogram of an original clean speech uttered by a male speaker and that of signals corrupted by white, train and babble noises at an SNR of 0 dB.

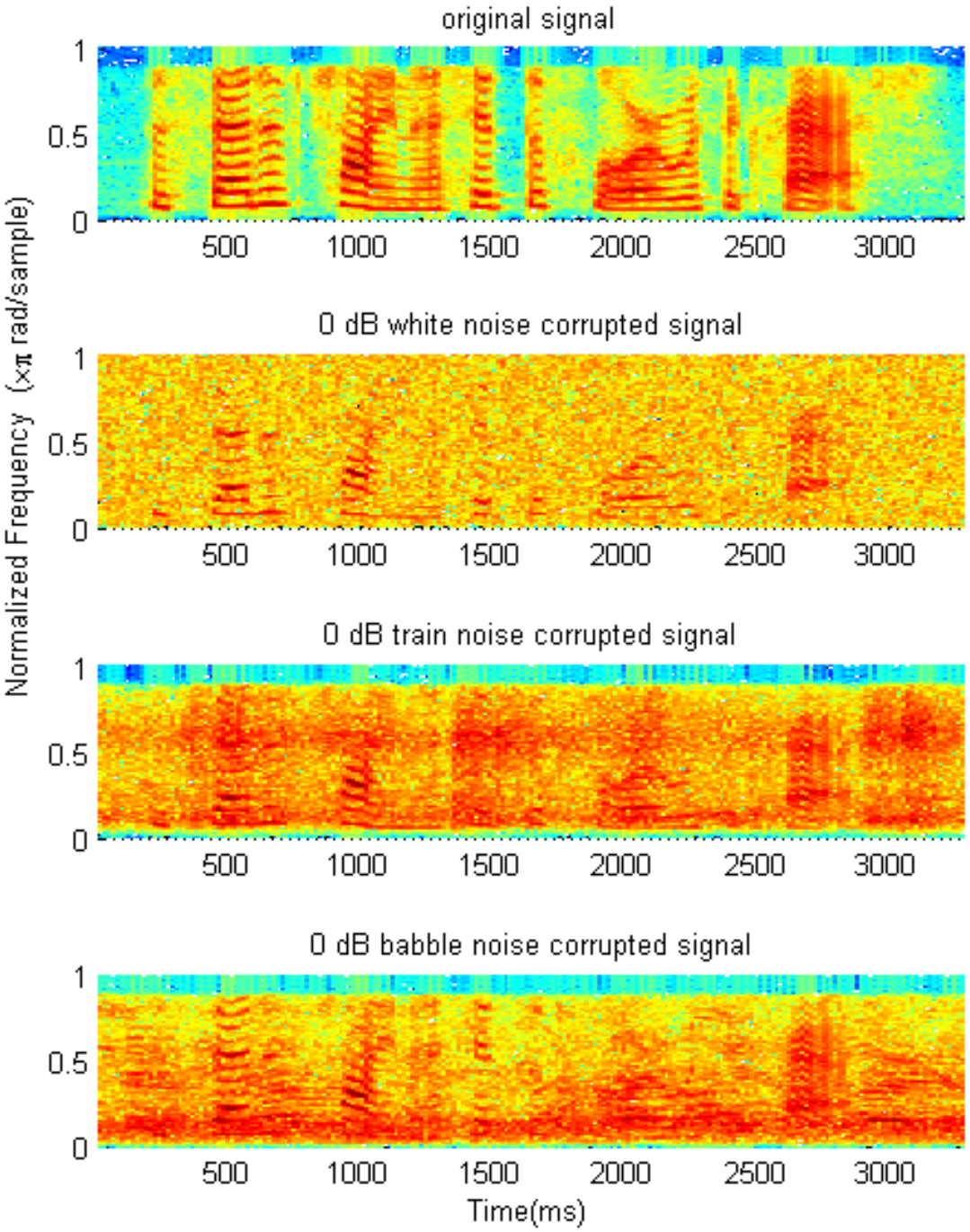

Fig. 4.8: Spectrogram of an original clean speech uttered by a female speaker and that of signals corrupted by white, train and babble noises at an SNR of 5 dB.

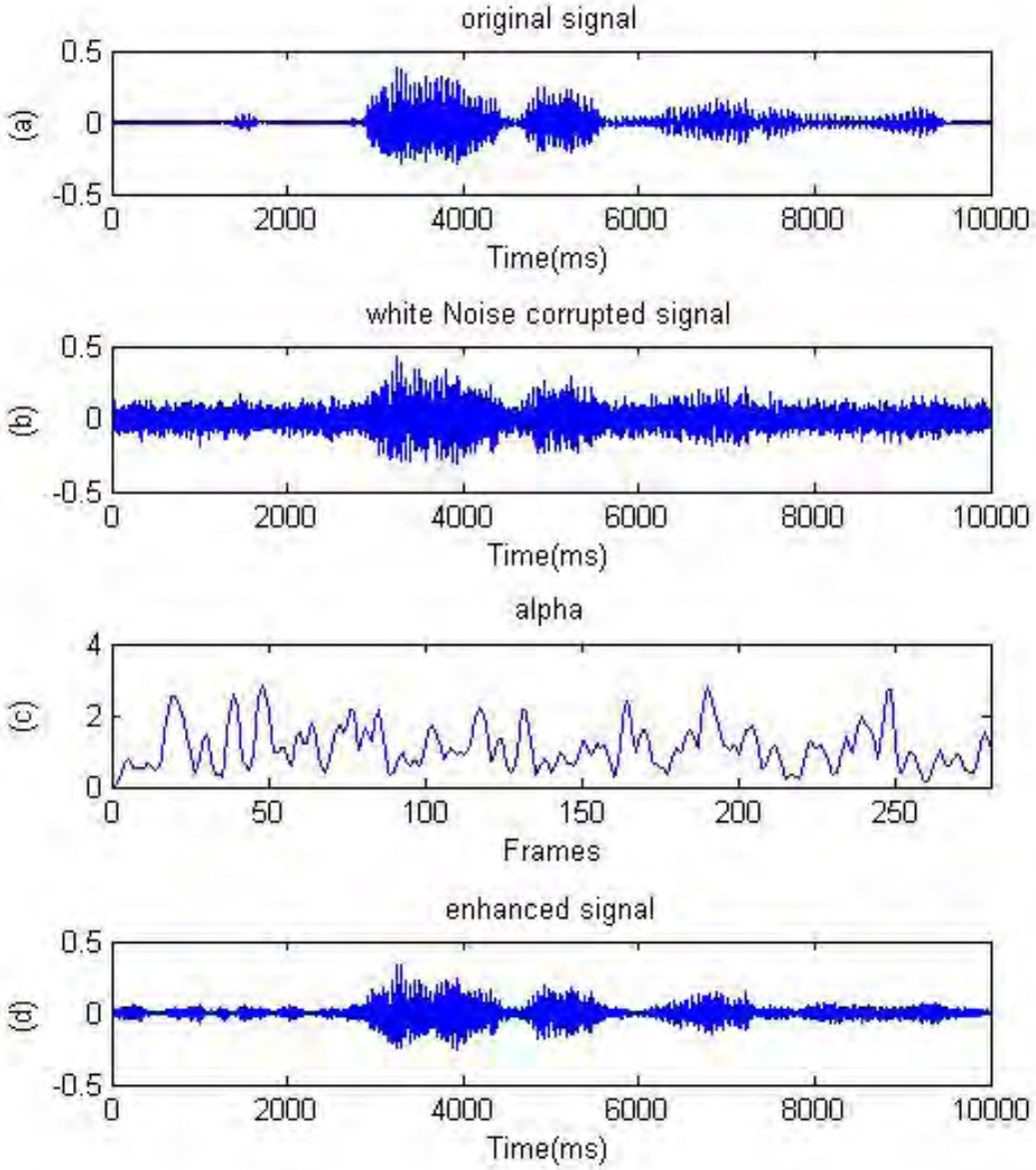

Fig. 4.9: (a) Original clean speech waveform. (b) White noise corrupted waveform at 10 dB. (c) Behavior of α_t over different frames (d) Waveform of enhanced speech obtained by using proposed method

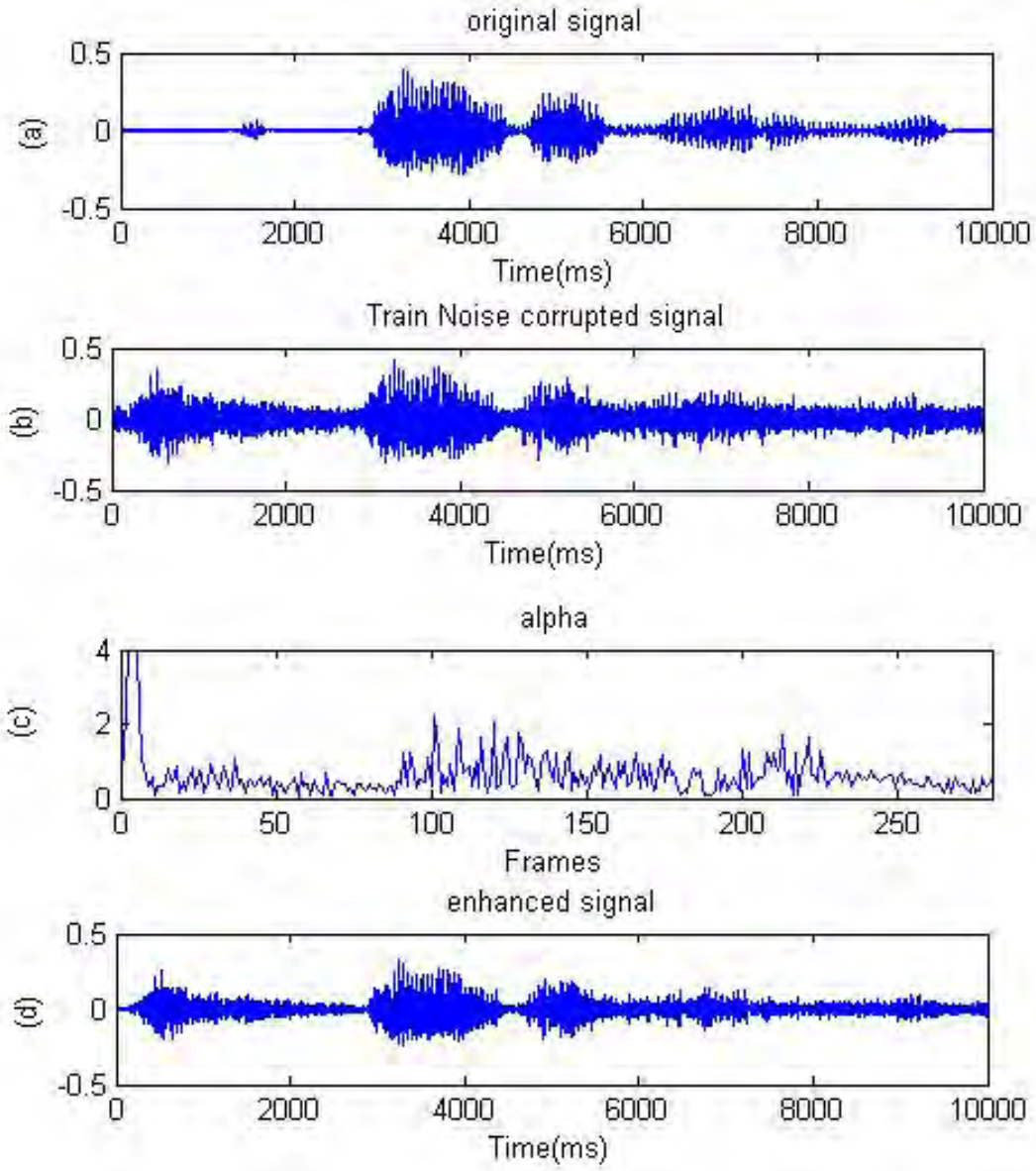

Fig. 4.10: (a) Original clean speech waveform. (b) Train noise corrupted waveform at 10 dB. (c) Behavior of α_t over different frames (d) Waveform of enhanced speech obtained by using proposed method

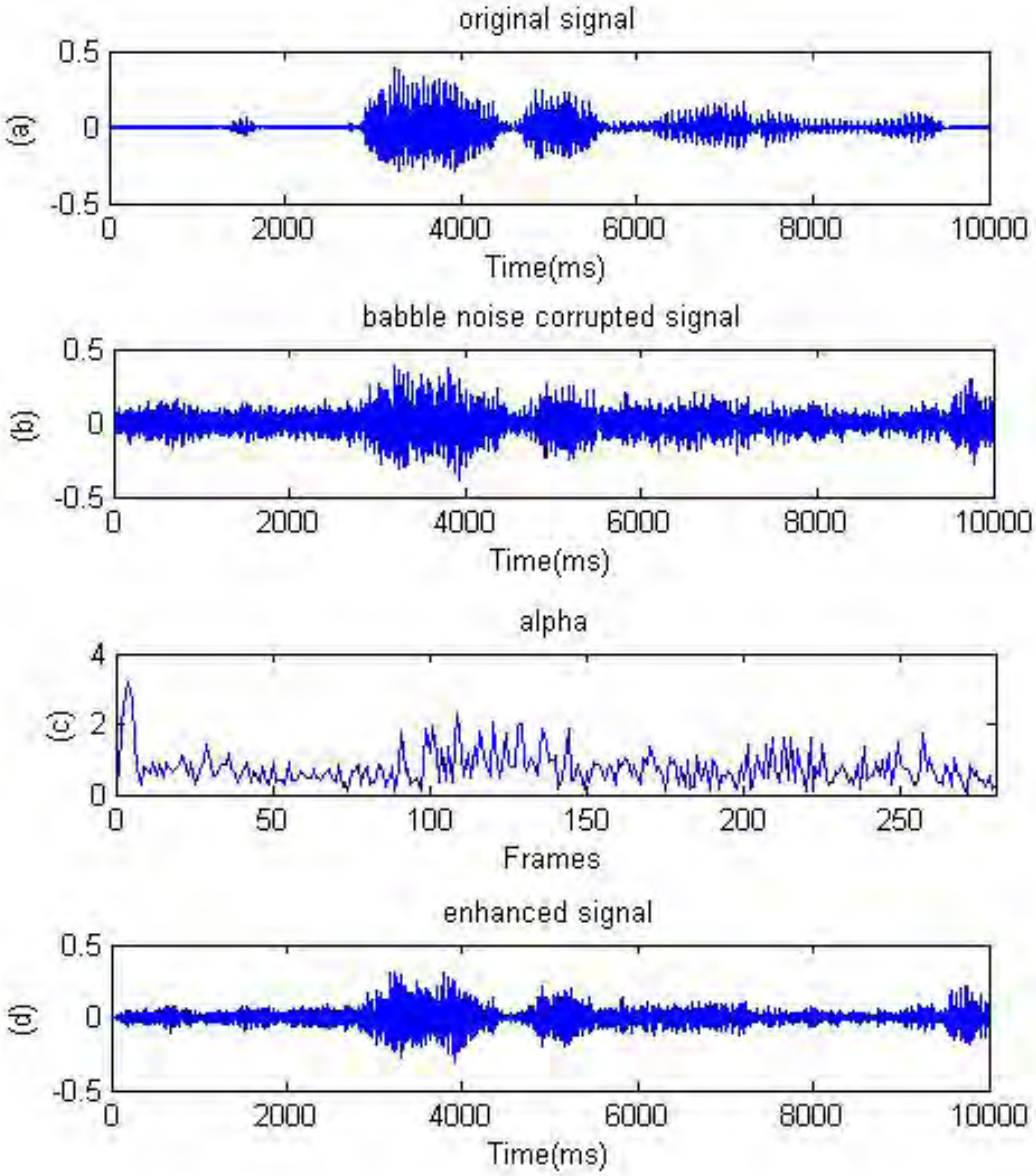

Fig. 4.11: (a) Original clean speech waveform. (b) Babble noise corrupted waveform at 10 dB. (c) Behavior of α_t over different frames (d) Waveform of enhanced speech obtained by using proposed method

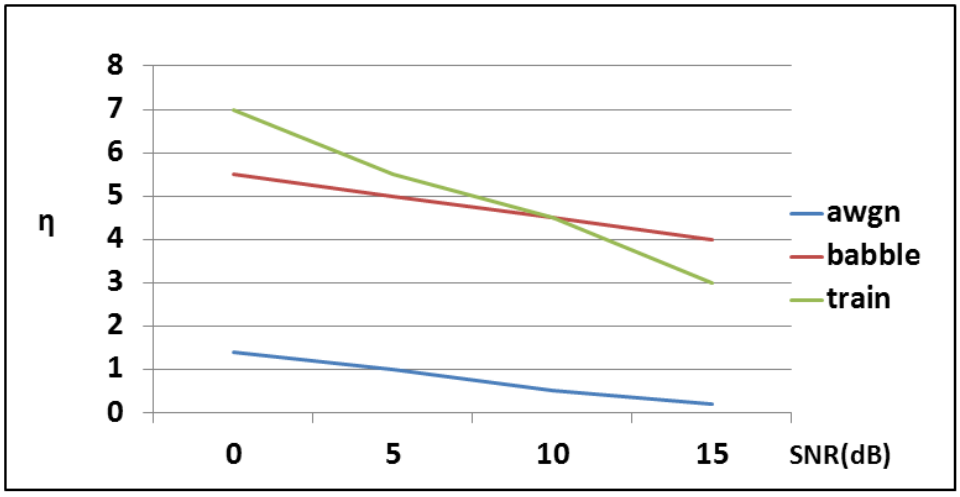

Fig. 4.12: Empirically determined η as a function of input speech SNR

The segmental SNR employed here is the average of the SNRs evaluated from all the segments or frames. The improvement in segmental SNR is given by

$$SegSNR_{imp} = SegSNR_{output} - SegSNR_{input}, \tag{4.2}$$

where $SegSNR_{input}$ and $SegSNR_{output}$ are the segmental SNRs of the input and output speech signals, respectively. The PESQ score calculation [45] is a hybrid of two perpetually motivated objective speech quality measures, such as PAMS and PSQM99. PESQ score maps means opinion scores(MOS) to a range between -0.5 and 4.5, where 1.0 corresponds to bad and 4.5 corresponds to distortion less. The proposed method is also subjectively evaluated in terms of informal listening test [44], [25], [46] and that the spectrogram representation of clean speech , noisy speech and enhanced speech. The performance of our method is compared with some of the existing method, such as conventional Boll's spectral subtraction(SSUB) [10], Shimamura's spectral subtraction [11], and paliwal's conjugate symmetry(ECSM) based short-time Fourier spectrum [23] in both objective and subjective senses.

4.1.2 Objective Evaluations

(1)Results on White Noise Corrupted Signal

For white noise-corrupted speech, Figs. 4.13 through 4.16 plots the segmental SNR improvement in dB for different methods while using a subset consists of eight speech sentences of the NOIZEOUS database at various SNRs (15 dB to 0 dB). From these

plots, it is clear that segmental SNR improvement in dB is higher with the proposed method for all the sentences at a wide range SNR considered. The results in terms of the objective metric, such as segmental SNR improvement in dB obtained by using the Boll's, Shimmamura's, Paliwal's and proposed methods for white noise corrupted speech are summarized and presented in Fig. 4.17. Fig. 4.17 shows the mean segmental SNR improvement in dB obtained by using different methods in the presence of white noise, where the SNR varies from 15 dB to 0 dB. It is seen from the Figure that in the SNR range under consideration, the comparison methods show comparatively lower values for the improvement in segmental SNR in dB relative to the proposed method.

Again, Figs. 4.18 through 4.21 shows the PESQ improvement for different methods while using a subset consists of eight speech sentences of the NOIZEOUS database at various SNRs (15 dB to 0 dB). From these plots, it is clear that improvement in PESQ is higher with the proposed method for all the sentences at a wide range SNR considered. The results in terms of the objective metrics, such as PESQ improvement obtained by using the Boll's, Shimmamura's, Paliwal's and proposed methods for white noise corrupted speech are presented in Fig 4.22. The PESQ improvement as a function of SNR resulting from the different methods in presence of white noise are portrayed in Fig. 4.22. It is vivid from this Figure that the proposed method yields higher PESQ improvement at high to low SNR levels compared to the other methods. Thus the proposed method is shown to be capable of producing enhanced space with better quality, whereas the PESQ improvement resulting from the other methods are relatively lower even at a higher SNR of 15 dB.

(2)Results on Train and Babble Noise Corrupted Speech Signals

Now, we present the results in terms of the objective metrics as mentioned above obtained by using all the methods in table I for train noise corrupted speech at SNR of 0dB. It is seen from the table I that at a particular SNR of 0 dB, the proposed method is superior in a sense that it produces the highest segmental SNR improvement in dB, whereas the other methods provide comparatively lower improvement. The performance of proposed method are also compared in table I in terms of PESQ improvement. It is observable that the PESQ improvement is the highest in case

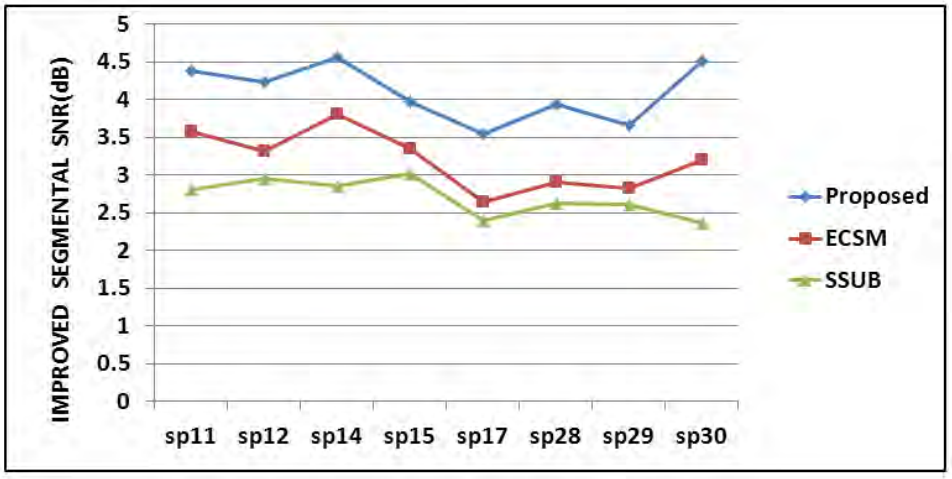

Fig. 4.13: Segmental SNR improvement comparing proposed method, ECSM method and SSUB method in case of white noise corrupted signal at SNR of 15dB for a subset consists of eight speech sentences of NOIZEOUS database.

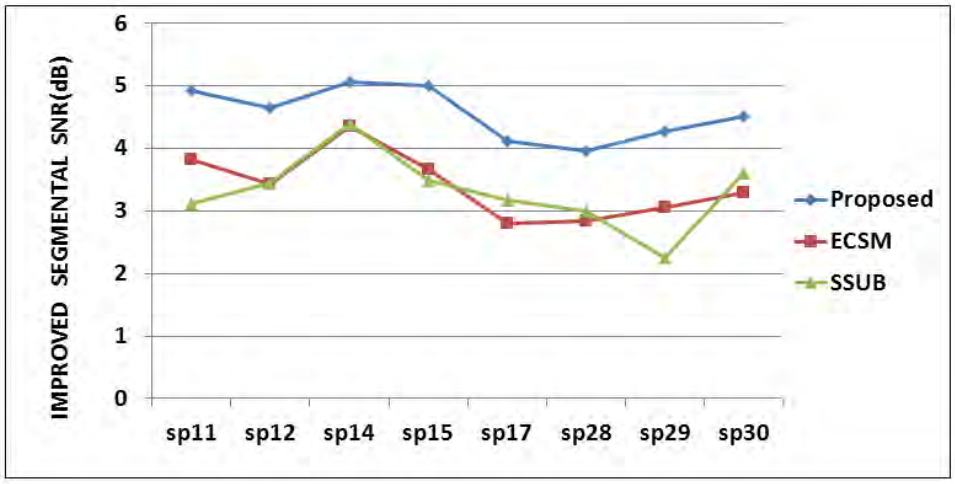

Fig. 4.14: Segmental SNR improvement comparing proposed method, ECSM method and SSUB method in case of white noise corrupted signal at SNR of 10 dB for a subset consists of eight speech sentences of NOIZEOUS database.

Fig. 4.15: Segmental SNR improvement comparing proposed method, ECSM method and SSUB method in case of white noise corrupted signal at SNR of 5dB for a subset consists of eight speech sentences of NOIZEOUS database.

Fig. 4.16: Segmental SNR improvement comparing proposed method, ECSM method and SSUB method in case of white noise corrupted signal at SNR of 0 dB for a subset consists of eight speech sentences of NOIZEOUS database.

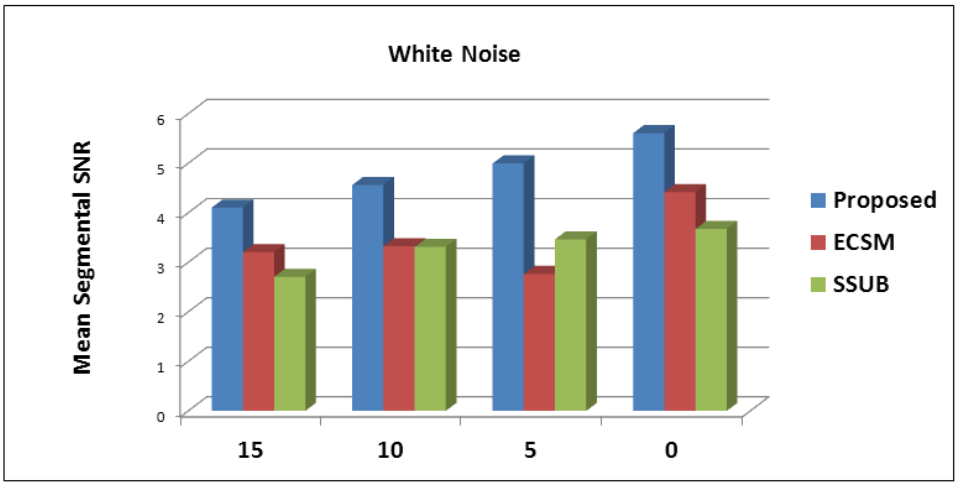

Fig. 4.17: Mean segmental SNR improvement comparing proposed method, ECSM method and SSUB method in case of white noise corrupted speech signal at SNRs of 15 dB, 10 dB, 05 dB and 0 dB.

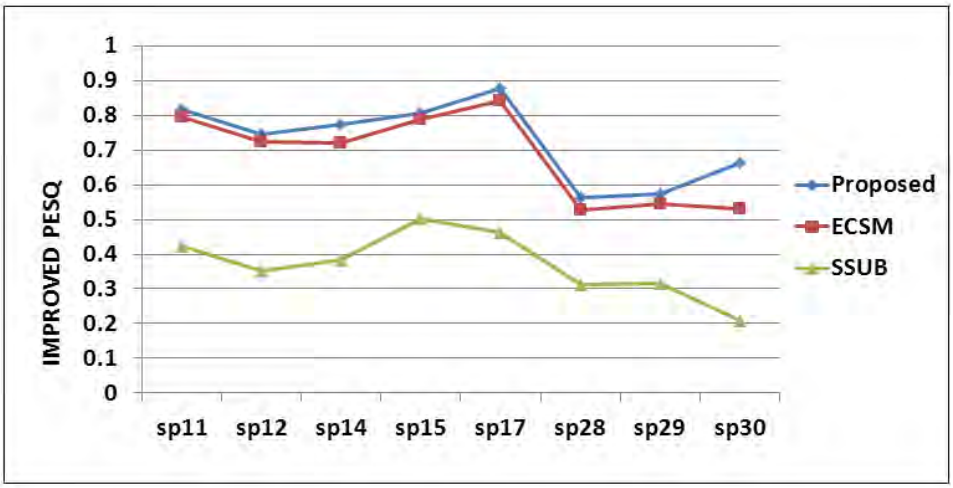

Fig. 4.18: PESQ improvement comparing proposed method, ECSM method and SSUB method in case of white noise corrupted signal at SNR of 15 dB for a subset consists of eight speech sentences of NOIZEOUS database.

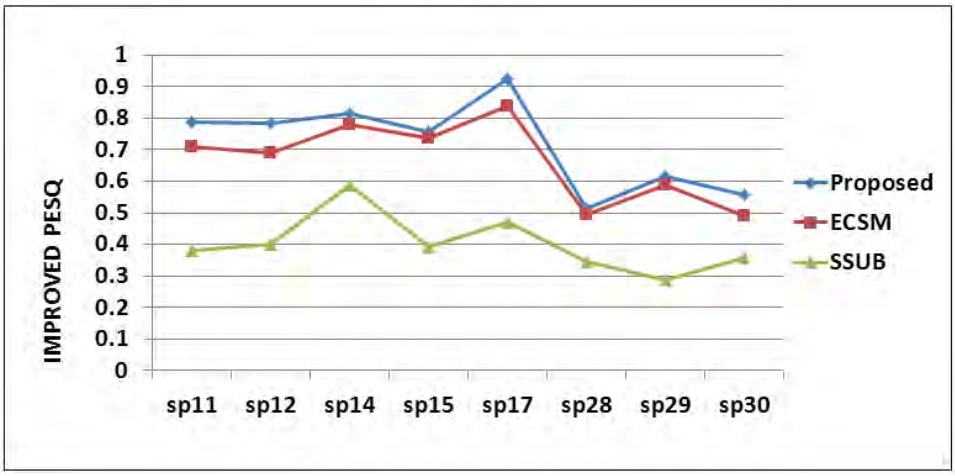

Fig. 4.19: PESQ improvement comparing proposed method, ECSM method and SSUB method in case of white noise corrupted signal at SNR of 10 dB for a subset consists of eight speech sentences of NOIZEOUS database.

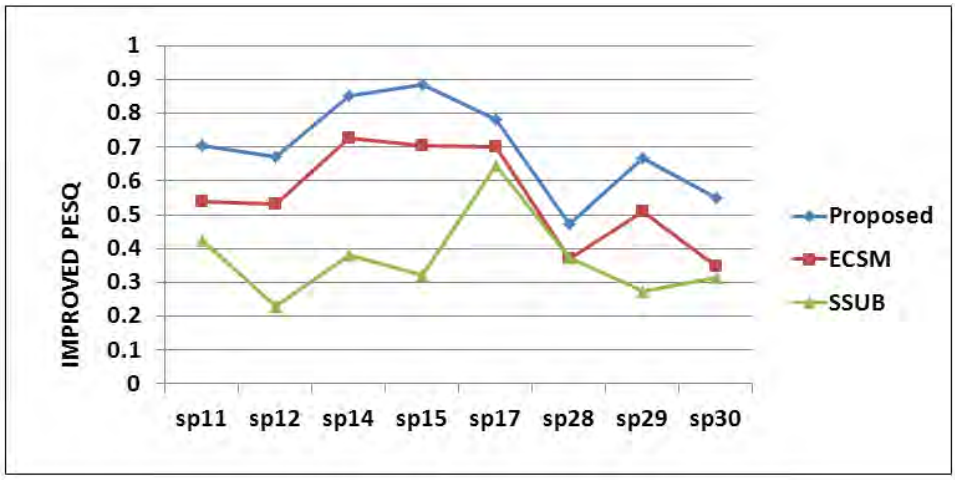

Fig. 4.20: PESQ improvement comparing proposed method, ECSM method and SSUB method in case of white noise corrupted signal at SNR of 5 dB for a subset consists of eight speech sentences of NOIZEOUS database.

Fig. 4.21: PESQ improvement comparing proposed method, ECSM method and SSUB method in case of white noise corrupted signal at SNR of 0 dB for a subset consists of eight speech sentences of NOIZEOUS database.

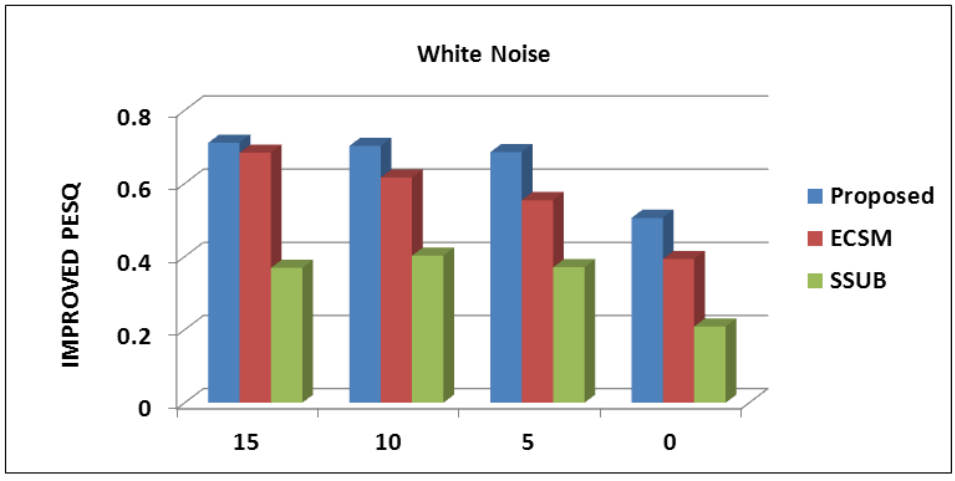

Fig. 4.22: Mean PESQ improvement comparing proposed method, ECSM method and SSUB method in case of white noise corrupted speech signal at SNRs of 15 dB, 10 dB, 05 dB and 0 dB.

Table 4.1: Performance Comparison in train noise at 0 db

Train Noise			
Metrics	ECSM	SSUB	Proposed
Segmental SNR Improvement	3.170206	2.72114	3.24454
PESQ Improvement	0.4688553	0.273646	0.5238999

Table 4.2: Performance Comparison in babble noise at 0 db

Babble Noise			
Metrics	ECSM	SSUB	Proposed
Segmental SNR Improvement	2.591334	1.784758	2.856127
PESQ Improvement	0.4131465	0.195114	0.4447725

of proposed method. Since a higher improvement indicates a better speech quality, the proposed method is indeed better in performance even in the presence of train noise.

In the presence of babble noise, the segmental SNR improvement in dB and PESQ improvement obtained by using the other methods are compared with respect to the proposed method and summmarized in table II at an SNR 0 dB. As seen from the table that the proposed method still gives higher values of segmental SNR improvement in dB compared to other methods at the low level of SNR, such as 0 dB. It is clearly demonstrated from the table II that while the other methods continue to produce lower PESQ improvement, the proposed method remains better even at the SNR as low as 0 dB of babble noise. It is noticeable that the performance of all the methods degrade in the presence of babble noise compared to that in the train or white noise, but the proposed method retains its superiority with respect to all other methods in terms of the objective metrics under consideration.

4.1.3 Subjective Evaluations

In order to evaluate the subjective observation of the enhanced speech obtained by using the proposed method, spectrograms of the clean speech, the noisy speech

and enhanced speech signals obtained by using all the methods are presented in Fig. 4.23 through Fig. 4.25 for white, train and babble noise corrupted speech signals, respectively. The results are plotted for SNR level of 10 dB. It is evidient that harmonics are well preserved and amount of distortion is greatly reduced in the proposed method compared to the other methods, no matter the speech is corrupted by white or train or babble noise. Thus spectrogram observations with lower distortion even in the presence of babble noise also validate our claim of better speech quality as obtained in our objective evaluation in terms of higher speech SNR improvement in dB, and higher PESQ improvement relative to the other comparison methods.

Informal [44] listening tests are conducted by allowing the listeners to perceptually evaluate the clean, noisy and enhanced speech signals. It is found that the subjective sound quality resulting by the proposed method possess the highest correlation with that obtained from the objective evaluation while compared with the other methods in case of all noises considered at different levels of SNR.

4.2 Conclusion

In this chapter, the simulation results in term of objective metrics, spectrogram representation and informal subjective listening tests demonstrate that the proposed method is capable of enhancing speech in different noisy conditions with a better quality and less distortion in comparison of some of the existing methods in the literature.

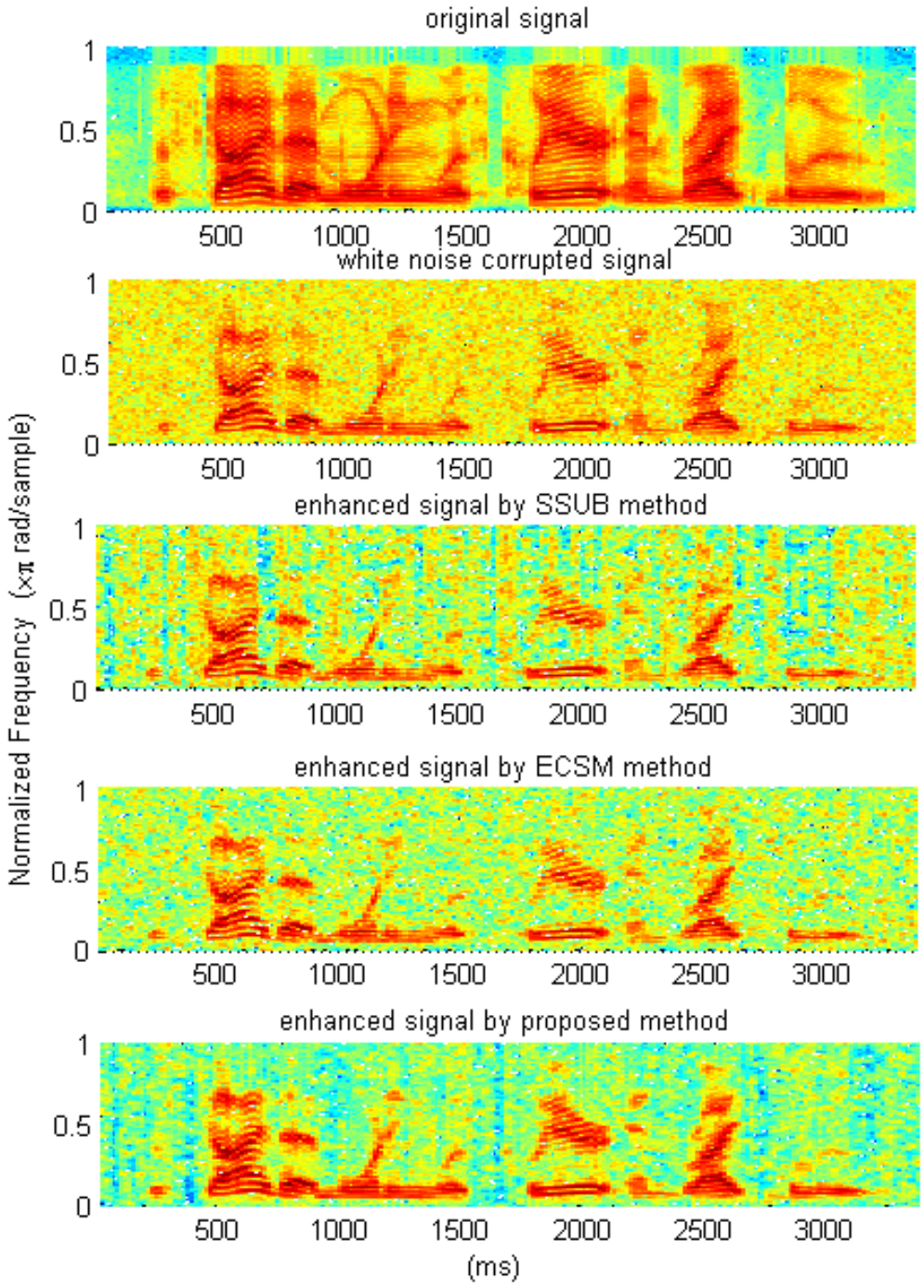

Fig. 4.23: Spectrograms of clean speech, noisy speech, and enhanced speech obtained by using the other and proposed methods for white noise at an SNR of 10 dB.

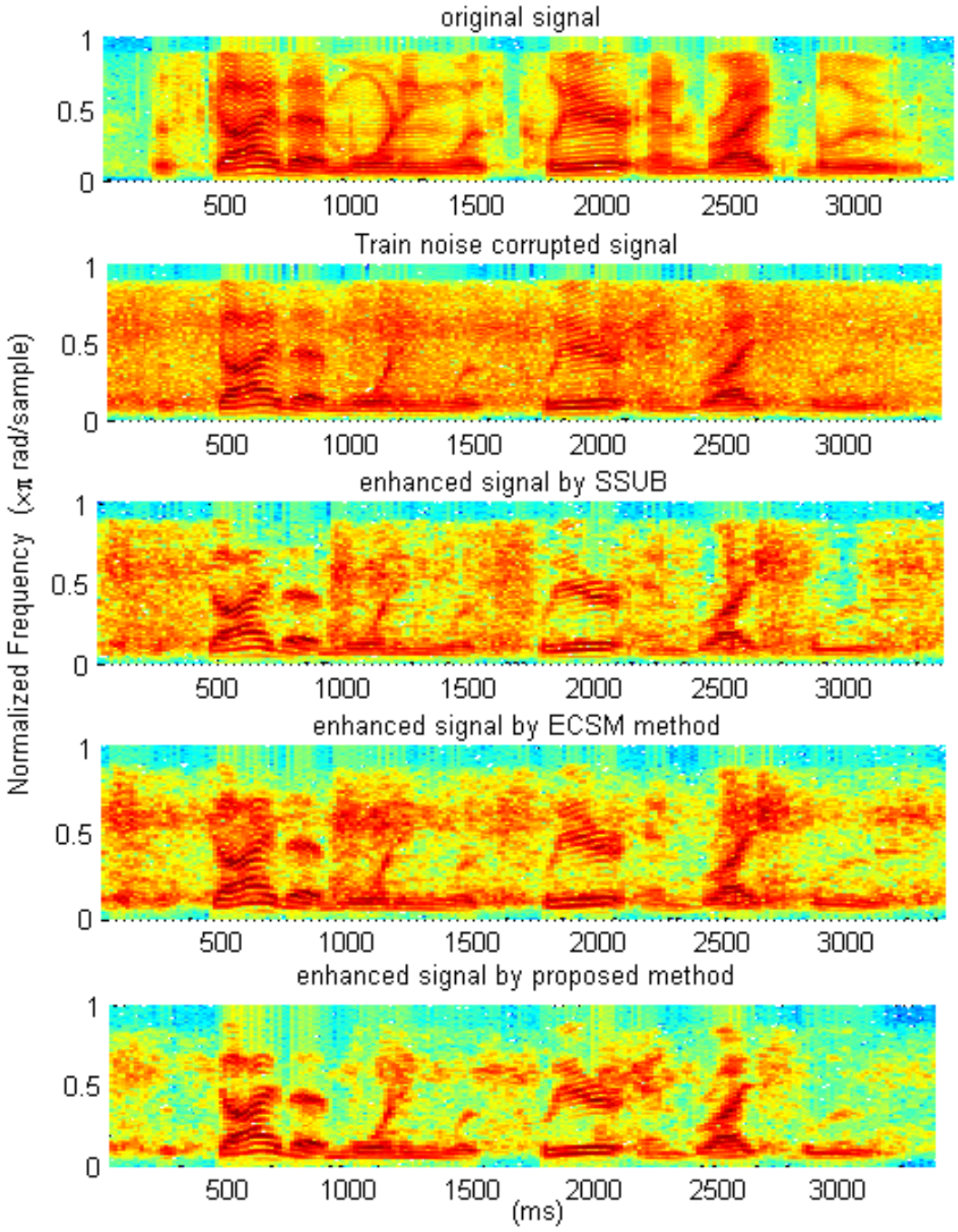

Fig. 4.24: Spectrograms of clean speech, noisy speech, and enhanced speech obtained by using the other and proposed methods for train noise at an SNR of 10 dB.

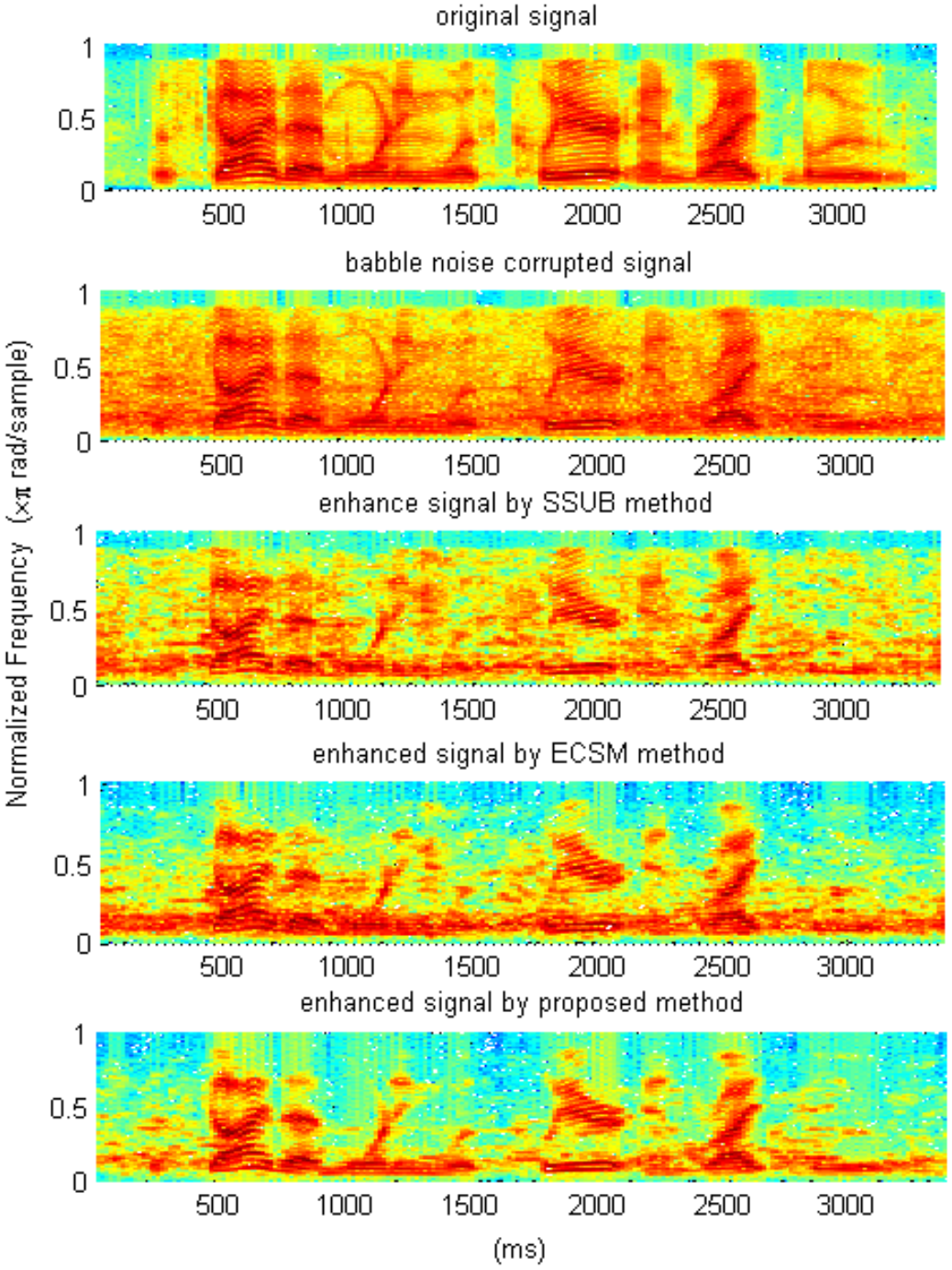

Fig. 4.25: Spectrograms of clean speech, noisy speech, and enhanced speech obtained by using the other and proposed methods for babble noise at an SNR of 10 dB.

Chapter 5

CONCLUSION

5.1 Concluding Remarks

In this thesis, unlike the conventional spectral subtraction method, a noisy speech enhancement method is developed based on noise compensation performed on short time magnitude as well phase spectra. Here the noise estimate is obtained by exploiting the low frequency regions of noisy speech of the current frame rather than depending only on the previous initial silence frames. We argue that such noise estimate can be used in a spectral subtraction approach to obtain a noise compensated magnitude spectrum. By employing the noise estimates thus obtained that offers the capability of tracking the time variation of the non-stationary noise, a procedure is formulated to compensate the distortion in the phase spectrum that is kept unchanged in the typical speech enhancement methods. The noise compensated phase spectrum is then recombined with the noise compensated magnitude spectrum to produce a modified complex spectrum thus synthesizing an enhanced frame.

5.2 Contribution of this Thesis

The major contributions of this thesis are,

1. Development of a single channel speech enhancement method based on the spectral subtraction, where compensations for the distortion in the magnitude and phase spectra are introduced.

2. Speech processed by the new method shows high levels of stationary and non-stationary noise suppression It is shown that noise compensations in magnitude

and phase of the noisy speech help preserving the speech content and improve the speech quality.

3. The modified spectral subtraction method along with phase correction provides a perceived improvement over the conventional methods and suffers minimally from musical noise.

4. Detail simulations are carried out in the presence of white train and babble noises from high to low lvel of SNRs to evaluate the performance of the proposed method in terms of standard objective metrics, namely segmental SNR improvement in dB and PESQ improvement as well as subjective evaluations, such as spectrogram representation and informal listening tests.

5. the performance of our method is compared with some of the existing methods, such as Boll's spectral subtraction(SSUB) [10], Shimamura's spectral subtraction [11], and paliwal's conjugate symmetry(ECSM) based short-time Fourier spectrum [23] in both objective and subjective senses.

6. Simulation results show that the proposed method yields consistently better results in terms of higher segmental SNR improvement and PESQ improvement than those of some of the existing methods and results in an enhanced speech with better qualities as well as intelligibility.

5.3 Scopes for Future Work

Further research can be conducted to address various issues which are still inherent to single channel subtractive type algorithms and issues involved in implementing the system in real-time.

1. In the proposed enhancement scheme, the noise compensations in magnitude and phase spectra can also be explored in other transform domains, such as discrete cosine transform(DCT) and discrete wavelet transform(DWT)

2. In the proposed speech enhancement scheme, noise compensations in the magnitude and phase spectra can be employed by using a perceptually weighted filter that would be able to mask the remaining residual noise making it audibly imperceptible.

3. In general speech quality and intelligibility vary for different speech sounds. This is due to the fact that noise has a non-uniform effect on various classes of phonemes [selective magnitude subtraction]. For the proposed algorithm as well, the performance varies for different speech sections. In general, the loss of quality and intelligibility is greater in low energy sections and during transitional segments compared to strong voiced segments like vowels. Therefore, the noise suppression rules need to be somehow modified based on speech classification, and get a decision on the tradeoff between noise suppression and speech distortion more locally.This can be also approached by incorporating a non-linear approach for subtraction within a frame, where subtraction parameters are different within a frame and depend on factors like frequency and phase ,etc.

Bibliography

[1] P. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 857–869, 2005.

[2] ——, *Speech enhancement: theory and practice*. CRC, 2007, vol. 30.

[3] J. Deller, J. Proakis, and J. Hansen, "Discrete-time processing of speech signals." Institute of Electrical and Electronics Engineers, 2000.

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

[5] N. Wiener, *Extrapolation, interpolation and smoothing of stationary time series. With engineering applications*. New York, 1950.

[6] N. Ma, M. Bouchard, and R. Goubran, "Speech enhancement using a masking threshold constrained kalman filter and its heuristic implementations," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 19–32, 2006.

[7] J. Allen and L. Rabiner, "A unified approach to short-time fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[8] R. Crochiere, "A weighted overlap-add method of short-time fourier analysis/synthesis," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 1, pp. 99–102, 1980.

[9] M. Portnoff, "Short-time fourier analysis of sampled speech," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 3, pp. 364–373, 1981.

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

[11] K. Yamashita and T. Shimamura, "Nonstationary noise estimation using low-frequency regions for spectral subtraction," *Signal Processing Letters, IEEE*, vol. 12, no. 6, pp. 465–468, 2005.

[12] B. Chen and P. Loizou, "A laplacian-based mmse estimator for speech enhancement," *Speech communication*, vol. 49, no. 2, pp. 134–143, 2007.

[13] R. Martin, "Spectral subtraction based on minimum statistics," *power*, vol. 6, p. 8, 1994.

[14] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 700–708, 2003.

[15] C. You, S. Koh, and S. Rahardja, "An invertible frequency eigendomain transformation for masking-based subspace speech enhancement," *Signal Processing Letters, IEEE*, vol. 12, no. 6, pp. 461–464, 2005.

[16] J. Chang, "Warped discrete cosine transform-based noisy speech enhancement," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 52, no. 9, pp. 535–539, 2005.

[17] Y. Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 1, pp. 59–67, 2004.

[18] J. Yamauchi and T. Shimamura, "Noise estimation using high frequency regions for spectral subtraction," *Ieice Transactions On Fundamentals Of Electronics Communications And Computer Sciences E Series A*, vol. 85, no. 3, pp. 723–727, 2002.

[19] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 8, pp. 799–807, 2001.

[20] M. Hasan, S. Salahuddin, and M. Khan, "A modified a priori snr for speech enhancement using spectral subtraction rules," *Signal Processing Letters, IEEE*, vol. 11, no. 4, pp. 450–453, 2004.

[21] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 845–856, 2005.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985.

[23] K. Wójcicki, M. Milacic, A. Stark, J. Lyons, and K. Paliwal, "Exploiting conjugate symmetry of the short-time fourier spectrum for speech enhancement," *Signal Processing Letters, IEEE*, vol. 15, pp. 461–464, 2008.

[24] S. So, K. Wocicki, J. Lyons, A. Stark, and K. Paliwal, "Kalman fitler with phase spectrum compensation algorithm for speech enhancement," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on.* IEEE, 2009, pp. 4405–4408.

[25] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[26] M. Bhatnagar, "A modified spectral subtraction method combined with perceptual weighting for speech enhancement," Master's thesis, UT Dallas, 2002.

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

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

[29] Y. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *Acoustics, Speech, and Signal Pro-*

cessing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on, vol. 2. IEEE, 2001, pp. 737–740.

[30] S. Ogata and T. Shimamura, "Reinforced spectral subtraction method to enhance speech signal," in *Electrical and Electronic Technology, 2001. TENCON. Proceedings of IEEE Region 10 International Conference on*, vol. 1. IEEE, 2001, pp. 242–245.

[31] W. Kim, S. Kang, and H. Ko, "Spectral subtraction based on phonetic dependency and masking effects," in *Vision, Image and Signal Processing, IEE Proceedings-*, vol. 147, no. 5. IET, 2000, pp. 423–427.

[32] G. Kang and L. Fransen, "Quality improvement of lpc-processed noisy speech by using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 6, pp. 939–942, 1989.

[33] S. Chang, Y. Kwon, S. Yang, I. Kim *et al.*, "Speech enhancement for non-stationary noise environment by adaptive wavelet packet," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–561.

[34] K. Paliwal, "Estimation of noise variance from the noisy ar signal and its application in speech enhancement," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 2, pp. 292–294, 1988.

[35] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[36] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and wiener filtering," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1875–1878.

[37] B. Widrow and S. Stearns, "Adaptive signal processing," *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1985, 491 p.*, vol. 1, 1985.

[38] W. Verhelst and O. Steenhaut, "A new model for the short-time complex cepstrum of voiced speech," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 1, pp. 43–51, 1986.

[39] S. Young, "A review of large-vocabulary continuous-speech," *Signal Processing Magazine, IEEE*, vol. 13, no. 5, p. 45, 1996.

[40] H. Kim and R. Rose, "Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for asr in noisy environments," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 435–446, 2003.

[41] Asaduzzaman and C. Shahnaz., "A spectral enhancement method for noisy speech based on noise compensated magnitude and phase spectra," *Submitted to international Journal of speech Technology, Springer.*

[42] A. Stark, K. Wójcicki, J. Lyons, and K. Paliwal, "Noise driven short-time phase spectrum compensation procedure for speech enhancement," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[43] E. Rothauser, W. Chapman, N. Guttman, K. Nordby, H. Silbiger, G. Urbanek, and M. Weinstock, "Ieee recommended practice for speech quality measurements," *IEEE Transactions on Audio Electroacoustics*, vol. 17, pp. 227–246, 1969.

[44] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[45] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. Ieee, 2001, pp. 749–752.

[46] D. O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965–2979, 2008.