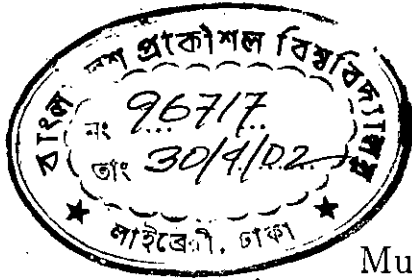


# Speech Enhancement by Combined Application of Hard and Soft Thresholding with Bias-compensated Noise Level

A thesis submitted to the Department of Electrical and Electronic Engineering  
of  
Bangladesh University of Engineering and Technology  
in partial fulfillment of the requirements for the degree of  
MASTER OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING



by

Muhammad Shamsul Arefeen Zilany

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING  
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

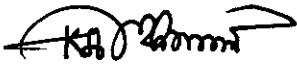
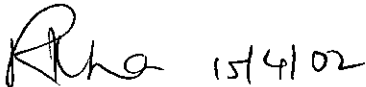
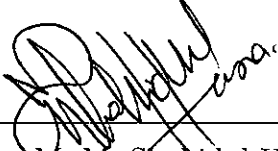

April 2002



#96717#

The thesis entitled “Speech Enhancement by Combined Application of Hard and Soft Thresholding with Bias-compensated Noise Level” submitted by Muhammad Shamsul Arefeen Zilany, Roll No.: 040006217p, Session: April, 2000 has been accepted as satisfactory in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING.

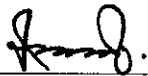
**BOARD OF EXAMINERS**

1.   
\_\_\_\_\_  
(Dr. Md. Kamrul Hasan )  
Associate Professor  
Department of Electrical and  
Electronic Engineering, BUET  
Dhaka-1000, Bangladesh. **Chairman**
  
2.   
\_\_\_\_\_  
(Dr. M. Rezwan Khan )  
Professor  
Department of Electrical and  
Electronic Engineering, BUET  
Dhaka-1000, Bangladesh. **Member**
  
3.   
\_\_\_\_\_  
(Dr. M. M. Shahidul Hassan)  
Professor and Head  
Department of Electrical and  
Electronic Engineering, BUET  
Dhaka-1000, Bangladesh. **Member**  
(Ex-officio)
  
4.   
\_\_\_\_\_  
(Dr. Md. Abul Kashem Mia)  
Associate Professor and Head  
Department of Computer Science  
and Engineering, BUET  
Dhaka-1000, Bangladesh. **Member**  
(External)

# Declaration

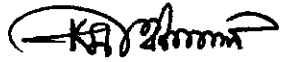
It is hereby declared that this thesis or any part of it has not been submitted elsewhere for the award of any degree or diploma.

Signature of the candidate



(Muhammad Shamsul Arefeen Zilany)

Signature of the supervisor



(Dr. Md. Kamrul Hasan )

Associate Professor

Department of Electrical and Electronic Engineering, BUET

Dhaka-1000, Bangladesh

# Acknowledgements

I would like to thank Dr. Md. Kamrul Hasan, Associate Professor, Dept. of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Bangladesh, for his friendly supervision, constructive suggestions and constant support during this research. I also appreciate his nagging and chiding me for enhancing my research effort.

I'm also indebted to Professor M. Rezwana Khan for his insightful comments and advice whenever required. I found him very cooperative in discussions of the problems during my research work. Thanks to Mohammad Imamul Hasan Bhuiyan, S. M. Mahbubur Rahman and others as they endowed me with their advocacy and encouragement for successful completion of this research.

Of course, I am grateful to my parents for their patience and love. I dedicate this thesis to my parents.

# Contents

<b>Acknowledgements</b>	<b>iii</b>
<b>List of Tables</b>	<b>vi</b>
<b>List of Figures</b>	<b>vii</b>
<b>Abstract</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Speech Enhancement: Background . . . . .	1
1.2 Objective of This Research . . . . .	5
1.3 Organization of the Thesis . . . . .	6
<b>2 Review of Speech Enhancement Techniques</b>	<b>8</b>
2.1 Introduction . . . . .	8
2.2 Speech Enhancement Techniques Based on Short-Time Spectral Amplitude Estimation . . . . .	8
2.2.1 Speech enhancement based on direct estimation of short- time spectral amplitude . . . . .	9
2.2.2 Speech enhancement techniques based on Wiener filtering .	12
2.3 Speech Enhancement Techniques Based on Speech Model . . . . .	13
2.4 Wavelet Speech Enhancement Based on the Teager Energy Operator	15
2.4.1 Wavelet packet analysis . . . . .	15
2.4.2 Teager energy approximation . . . . .	16
2.4.3 Masks construction . . . . .	16
2.4.4 Threshold modulation criterion . . . . .	16
2.4.5 Mask processing for the time-adapted threshold . . . . .	17
2.4.6 Time-adapted threshold . . . . .	17
2.4.7 Thresholding $WP$ coefficients . . . . .	17

2.4.8	Inverse transformation . . . . .	18
2.5	Conclusion . . . . .	18
<b>3</b>	<b>Bias-compensated Noise Level for Wavelet and DCT Speech Enhancement</b>	<b>19</b>
3.1	Introduction . . . . .	19
3.2	Problem Formulation . . . . .	20
3.3	Estimation of Noise Level: Conventional Approach . . . . .	21
3.4	Wavelet Transform Based Proposed Enhancement Algorithm . . .	21
3.4.1	Calculation of corrected value of noise level . . . . .	23
3.4.2	Thresholding <i>WP</i> coefficients . . . . .	27
3.4.3	Reconstruction of the original signal . . . . .	28
3.5	DCT Based Proposed Enhancement Algorithm . . . . .	29
3.5.1	Calculation of corrected value of noise level . . . . .	30
3.5.2	Thresholding DCT coefficients . . . . .	30
3.6	Conclusion . . . . .	32
<b>4</b>	<b>Results</b>	<b>33</b>
4.1	Data Used . . . . .	33
4.2	Estimation of Corrected Noise Level . . . . .	33
4.3	Performance Test . . . . .	35
4.3.1	Objective test . . . . .	38
4.3.2	Subjective test . . . . .	53
4.4	Conclusion . . . . .	55
<b>5</b>	<b>Conclusion</b>	<b>56</b>
5.1	Summary . . . . .	56
5.2	Suggestions for future work . . . . .	57
	<b>Bibliography</b>	<b>58</b>

# List of Tables

4.1	Comparison of actual and corrected noise levels along with the correction factor, $\beta$ , for the speech, “She had your dark suit in greasy wash water all year”, at different SNRs. . . . .	34
4.2	Comparison of actual and corrected noise levels along with the correction factor, $\beta$ , for the speech, “Should we chase those cowboys?”, at different SNRs. . . . .	36
4.3	Results on SNR improvement for the speech, “Should we chase those cowboys?”, corrupted by additive white noise . . . . .	38
4.4	Results on SNR improvement for two utterances (s3 and s4) corrupted by recorded real car noise . . . . .	40
4.5	Results of subjective evaluation in terms of preference percentage in wavelet domain . . . . .	54
4.6	Results of subjective evaluation in terms of preference percentage in DCT domain . . . . .	54

# List of Figures

2.1	The spectral subtraction approach . . . . .	10
2.2	A speech production model . . . . .	14
3.1	Variation of $\sigma_{v+}$ with SNR: (a) wavelet; (b) DCT. . . . .	22
3.2	Variation of scaled kurtosis with SNR in wavelet domain: (a) $\gamma(\zeta_p)$ vs. SNR; (b) Shifted $\gamma(\zeta_p)$ vs. SNR. . . . .	24
3.3	Variation of scaled kurtosis with SNR in DCT domain: (a) $\gamma(\zeta_p)$ vs. SNR; (b) Shifted $\gamma(\zeta_p)$ vs. SNR. . . . .	31
4.1	Estimation of SNR of noisy speech. . . . .	37
4.2	Performance comparison in terms of input-output SNR for the speech, “She had your dark suit in greasy wash water all year”, corrupted by additive white noise. . . . .	39
4.3	Enhancement results for a female utterance “She had your dark suit in greasy wash water all year” corrupted by additive white noise: (a) Time-domain; (b) Spectrogram; (i) clean, (ii) noisy, (iii) denoised using Ref. [37], (iv) soft thresholding in wavelet domain, (v) hard and soft thresholding in wavelet domain, (vi) soft thresholding in DCT domain, (vii) hard and soft thresholding in DCT domain. . . . .	43
4.4	Enhancement results for the utterance by a male speaker “Should we chase those cowboys?” corrupted by additive white noise: (a) Time-domain; (b) Spectrogram; (i) clean, (ii) noisy, (iii) denoised using Ref. [37], (iv) soft thresholding in wavelet domain, (v) hard and soft thresholding in wavelet domain, (vi) soft thresholding in DCT domain, (vii) hard and soft thresholding in DCT domain. . .	46

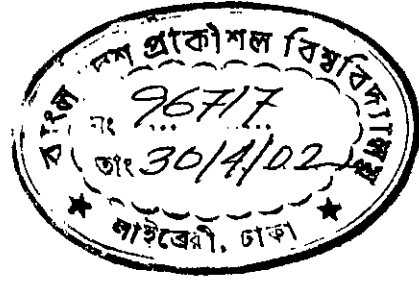


4.5	Enhancement results for the utterance by a male speaker “Would you please confirm government policy” corrupted by real car noise: (a) Time-domain; (b) Spectrogram; (i) clean, (ii) noisy, (iii) denoised using Ref. [37], (iv) soft thresholding in wavelet domain, (v) hard and soft thresholding in wavelet domain, (vi) soft thresholding in DCT domain, (vii) hard and soft thresholding in DCT domain. . . . .	49
4.6	Enhancement results for the numerical counting “One two” by a female speaker corrupted by the recorded real car noise: (a) Time-domain; (b) Spectrogram; (i) clean, (ii) noisy, (iii) denoised using Ref. [37], (iv) soft thresholding in wavelet domain, (v) hard and soft thresholding in wavelet domain, (vi) soft thresholding in DCT domain, (vii) hard and soft thresholding in DCT domain. . . . .	52

# Abstract

This thesis presents an improved thresholding technique with an accurate estimation of noise level for speech enhancement in both wavelet and DCT domain. Usually, setting of a threshold criteria requires an accurate estimation/knowledge of the noise level in the noisy speech. In speech, most of the signal energy is retained in the lower frequency range. The higher frequency region essentially contains noise from which noise level is usually estimated. The amount of signal present in this region may be insignificant at low SNR but is not negligible particularly at high SNR. Because of the presence of a small fraction of signal coefficients at the high frequency region, we get an estimate of noise level that suffers an upward bias. For this reason, most of the existing techniques show deteriorating performance especially at high SNR.

In this work, the behaviour of the normalized kurtosis of the noisy transform coefficients at the finest level is exploited for compensating this bias. Unlike other conventional techniques, we propose a novel approach for computing a correction factor to reduce the upward bias of the noise level obtained from the median absolute deviation (*MAD*) of the high frequency coefficients of the degraded speech. This signal-bias compensated noise level is then used as the threshold parameter which prevents the fall of SNR even when the SNR of the given noisy speech is high. The corrected noise level can also be used for estimation of SNR more accurately. In this thesis, successive application of hard and soft thresholding is proposed to devise an improved denoising method. Regions in the transform domain, where signal strength is less than that of noise, hard thresholding is applied by forcing the coefficients to be zero. This will eliminate a significant portion of noise from the regions of coefficients where noise dominates signal. After accomplishing hard thresholding, soft thresholding is applied over the rest of the regions to further reduce the noise level. The performance of the proposed algorithm is evaluated on speech corrupted by background white Gaussian noise and real noise recorded inside a moving car.



# Chapter 1

## Introduction

### 1.1 Speech Enhancement: Background

Speech enhancement is the term used to describe algorithms or devices whose purpose is to improve some perceptual aspects of speech for the human listener or to improve the speech signal so that it may be better exploited by other speech processing algorithms. Development and widespread deployment of digital communication systems during the last twenty years have brought increased attention to the role of speech enhancement in speech processing problems [1]-[6]. Speech enhancement algorithms have been applied to problems as diverse as correction of reverberation, pitch modification, rate modification, reconstruction of lost speech packets in digital networks, correction of so-called "hyperbaric" speech produced by deep-sea divers breathing a helium-oxygen mixture and correction of speech that has been distorted due to pathological problems of the speaker. However, noise reduction is probably the most important and most frequently encountered speech enhancement problem.

Speech enhancement attempts to improve the performance of voice communication systems when their input or output signal is corrupted by noise. The improvement is in the sense of minimizing the effects of noise on the performance of these systems. The need for enhancing speech signals arises in many situations in which the speech either originates from some noisy location or is affected by the noise over the channel or at the receiving end. Both digital and analog channels are possible, and communication can be either between people or with a machine. Hence speech enhancement is the problem of enhancing a given sample function of noisy speech signal, as well as the problem of enhancing the performance of

speech coding and recognition systems whose input signal is noisy [1]-[40]. Examples of important applications of speech enhancement include improving the performance of 1) cellular radio telephone systems, which usually suffer from background noise in the car as well as from channel noise; 2) pay phones located in noisy environments (e.g., airports); 3) air-ground communication systems in which the cockpit noise corrupts the pilot's speech; 4) teleconferencing systems where noise sources in one location may be broadcast to all other locations; 5) long distance communication over noisy radio channels; 6) paging systems located in noisy environments (e.g., airports, machine rooms); 7) ground-air communication in which the cockpit noise corrupts the received messages; and 8) suboptimal speech quantization systems.

In the cellular radio telephone example, the original speech is corrupted by the noise generated by the engine, fan, traffic and wind [7], [8], as well as by the channel noise. The signals delivered by cellular systems may therefore be noisy with impaired quality and intelligibility. If the cellular system encodes the signal prior to its transmission, then further degradation in its performance results, since speech coders rely on some model for the clean signal and normally that model is not suitable for the noisy signal. Similarly, if the cellular system is equipped with a speech recognition system which is used for automatic dialing, then the recognition accuracy of such system deteriorates in the presence of noise, since the noisy input signal is unlikely to obey the statistical model for the clean signal used by the recognizer. Similar problems are encountered with pay phone communication, air-ground communication, and teleconferencing systems. In the air-ground communication examples, however, the messages of low quality and intelligibility delivered to the air traffic controllers may have disastrous effects. The situation in long distance communication, paging systems, and ground-air communication is somewhat simpler, since the noise is added to the speech at the channel and at the receiving end, respectively, rather than at the source location. Hence, the clean signal can be "immunized" prior to being affected by the noise [9]-[11]. In suboptimal quantization of speech signals, the quantized signal is considered a noisy version of the clean signal [12]-[13]. Hence, enhancement can be applied to reduce the quantization noise, provided that quantization was not optimally performed.

The foregoing discussion demonstrates that speech enhancement has three major goals:

- 1) to improve perceptual aspects (e.g., quality, intelligibility) of a given sample function of degraded speech signal;
- 2) to increase robustness of speech coders to input noise;
- 3) to increase robustness of speech recognition systems to input noise.

The quality of speech signal is a subjective measure which reflects on the way the signal is perceived by listeners. It can be expressed in terms of how pleasant the signal sounds or how much effort is required on behalf of the listeners in order to understand the message. Intelligibility, on the other hand, is an objective measure of the amount of information which can be extracted by listeners from the given signal, whether the signal is clean or noisy. A given signal may be of high quality and low intelligibility, and vice versa. Hence, the two measures are independent of each other. Both the quality and the intelligibility of a set of given signals are evaluated based on tests performed on human listeners. Since no mathematical quantification of these measures, in terms of closed-form perceptually meaningful distortion measures, is known, algorithms for goals 1 and 2 above are difficult to design and evaluate. Goal 3 is significantly simpler since the problem is that of decoding the signal into a finite number of classes and the ultimate goal can be easily formulated in mathematical terms. Usually the problem is that of designing decoders which minimize the probability of recognition error.

The speech enhancement problem consists of a family of subproblems characterized by the type of noise source, the way the noise interacts with the clean signal, the number of voice channels, or microphone outputs, available for enhancement, and the nature of speech communication systems. The noise, or the interfering signals, may, for example be due to competitive speakers, background sounds (music, fans, machines, door slamming, wind, traffic, etc.), room reverberation, or random channel noise. The noise may accompany the original signal at the source location, over communication channels, or at the receiving end. It may affect the original signal in an additive, multiplicative, or convolutional manner. Furthermore, the noise may be statistically dependent or independent of the clean signal. The number of voice channels available for enhancement is

an important factor in designing speech enhancement systems. In general, the larger the number of microphones, the easier the speech enhancement task. The communication system for which speech enhancement is designed can simply be a recording which has to be displayed to audience, a man-machine communication system (speech recognizer), a digital communication system, etc.

Speech enhancement based on spectral decomposing and filtering [14]-[22] remains a common and effective approach for enhancing speech degraded by acoustic additive noise when only the noisy speech is available. This general class is based on variations of optimum filters and encompasses such methods as Wiener filtering, spectral subtraction and various maximum likelihood (ML) estimation schemes. A common set of requirements in this class includes: 1) An appropriate suppression rule based on an optimality criteria [15], [16] and usually function of the SNR (signal to noise ratio) and other speech and noise statistics. 2) An estimation of the speech and noise power spectral densities, or their respective autocorrelation. 3) A quantification of the probability of speech presence to further attenuate non-speech bands [17]. 4) A method for reducing residual noise by appropriately smoothing the estimated quantities [15] and/or exploiting the psychoacoustic properties of human hearing.

The choice of suppression rules is governed by many factors, such as computational efficiency, optimality criteria, and the exploiting of human hearing properties. In the reported literature, the range includes heuristic rules (e.g., [16]) as well as formally derived ones. The ML estimation approaches in [15], [18] attempt to better exploit the statistical properties of the DFT (discrete Fourier Transform) of noisy speech. These methods assume a statistical model for the DFT coefficients of noisy speech and derive optimum estimators of the magnitude spectrum based on that model.

An important contribution in this area is the smoothing approach proposed in [15] whereby the variation in SNR between successive frames is reduced by averaging the locally computed SNR ( $\text{SNR}_{post}$ ) with the SNR estimated in the previous frame after the filtering operation ( $\text{SNR}_{est}$ ). The method results in a significant reduction of the 'musical noise' artifacts, as shown in [14].

Another speech enhancement approach is the signal subspace (SS) method [23], [24]. The key idea is to decompose the vector space of the noisy signal into

a signal-plus-noise subspace and a noise subspace under the assumption that the additive noise is white. The enhancement is performed by removing the noise subspace and estimating the clean speech from the remaining signal-plus-noise subspace. Hidden Markov Model (HMM) based speech enhancement approaches [25], [26] have also drawn much attention in recent years.

Methods for speech enhancement have also been developed based on extraction of parameters from noisy speech, and synthesizing speech from these parameters [27]. All-pole modeling of degraded speech is one such method [28]. In all-pole modeling, if wrong peaks are extracted, then these peaks may get enhanced. Temporal sequence of these peaks also produces discontinuities in the contours of the spectral peaks when compared with the smooth contours in natural speech. Methods for speech enhancement have also been suggested based on the periodicity due to pitch [29]. Noise samples in successive glottal cycles are uncorrelated. On the other hand, the characteristics of the vocal tract system are highly correlated due to slow movement of the articulators. These methods for enhancement of speech depend critically on the estimation of pitch from the noisy speech signal.

Many speech enhancement algorithms make use of DFT to make it easier to remove noise embedded in the noisy speech signal [1]-[22]. Recently, discrete cosine transform (DCT) and wavelet transform have been widely used as analysis tools in the field of speech enhancement [30]-[36]. DCT is widely used because of its excellent energy compaction properties. During the past decade, wavelet transforms have been applied to various research areas which include signal and image denoising, compression, detection and pattern recognition. The application of wavelet shrinking for speech enhancement has been reported in many works [31], [32]. The wavelet transform combined with other signal processing tools has also been proposed for speech enhancement. Wiener filtering in the wavelet domain [33], wavelet filter bank for spectral subtraction [34] or coherence function [35], [36] are the examples of such methods.

## 1.2 Objective of This Research

The objective of this research is to extract the speech signal from the observed degraded version by applying a new speech enhancement technique based on the

combined application of hard and soft thresholding of the transform coefficients of the noisy signal with bias-compensated noise level as the threshold parameter. Unlike other conventional techniques, more accurate estimate of the threshold parameter, the noise level, is obtained by compensating the effect of the trace of the signal remaining at the high frequency region of the transform coefficients of the degraded speech. Using fourth-order statistics, we introduce a novel approach for computing a correction factor to reduce the upward bias of the noise level obtained from the median absolute deviation (*MAD*) of the high frequency coefficients of the degraded speech.

Here, we successively apply both hard and soft thresholding to devise an improved denoising method. The transform coefficients are first divided into a number of blocks consisting of convenient number of consecutive coefficients of the transformed signal. Then hard thresholding is applied to the blocks of coefficients where average signal power is less than average noise power as these blocks essentially contribute more noise than signal to the denoised speech. To identify the blocks for hard thresholding, a window of length same as block size is chosen and is slid over the whole range. This will eliminate a significant portion of noise from the regions of coefficients where noise dominates signal. The rest of the regions where signal strength is higher than that of noise, soft thresholding is applied for further enhancement of the noisy signal. As the noise power uniformly penetrates into the actual signal in the transformed domain, subtraction of noise power from the transformed noisy signal power is expected to improve the SNR of the enhanced signal. The coefficients with power less than the average noise power are more susceptible to distortion; their amplitudes are reduced proportionately.

We investigate performance of the proposed method in both wavelet and DCT (discrete cosine transform) domain using corrected noise level as the threshold parameter. The results will be compared with one of the most recent methods proposed by Bahoura and Rouat [37].

### 1.3 Organization of the Thesis

This thesis consists of five chapters. Chapter one gives an introduction followed by literature review and objective of the work.



In Chapter two a brief review of speech enhancement techniques are presented. These include brief illustration of traditional approaches such as spectral subtraction, Wiener filtering, enhancement based on speech model and wavelet based method.

In Chapter three a new approach for speech enhancement in wavelet and DCT domain considering the signal remaining at the finest level at the presence of additive white noise is investigated. The method for determining the bias-compensated noise level is extensively covered which is subsequently used as the threshold parameter. Both hard and soft thresholding criteria to be applied successively are then proposed.

The simulated speech enhancement results for the proposed method in both wavelet and DCT domain are presented in Chapter four in order to compare the results with the recent one proposed by Bahoura and Rouat [37]. Both subjective and objective evaluation are also reported along with necessary measurements in this chapter.

The thesis concludes by presenting an overall discussion on the work and pointing out some unsolved problems for future work in Chapter five.

# Chapter 2

## Review of Speech Enhancement Techniques

### 2.1 Introduction

Speech enhancement plays a key role in designing robust automatic speech and speaker recognition systems. As the presence of noise deteriorates the performance of the recognition systems and also shows an adverse effect on the perceived quality and intelligibility of speech at the receiving end, several approaches for speech enhancement in additive noise have been proposed. Speech enhancement based on spectral decomposition and variations of optimum filters cover the methods such as Wiener filtering, spectral subtraction and various maximum likelihood (ML) estimation schemes. Speech enhancement systems which can operate on the clean signal prior to its degradation by noise achieve significant improvement in the intelligibility of the noisy signal [9]-[11]. On the other hand, the systems which can operate on the signal only after it has been contaminated by noise primarily improve the quality of the noisy signal at the expense of some intelligibility loss [1], [2]. The major breakthrough in speech enhancement techniques are described in the following sections.

### 2.2 Speech Enhancement Techniques Based on Short-Time Spectral Amplitude Estimation

In general, in enhancement of a signal degraded by additive noise, it is significantly easier to estimate the spectral amplitude associated with the original signal than it is to estimate both amplitude and phase. It is principally the short-time

spectral amplitude rather than phase that is important for speech intelligibility and quality. There are a variety of speech enhancement techniques that capitalize on this aspect of speech perception by focusing on enhancing only the short-time spectral amplitude. The techniques to be discussed can be broadly classified into two groups. First, the short-time spectral amplitude is estimated in the frequency domain, using the spectrum of the degraded speech. Each short-time segment of the enhanced speech waveform in the time domain is then obtained by inverse transforming this spectral amplitude estimate combined with the phase of the degraded speech. In the second class, the degraded speech is first used to obtain a filter which is then applied to the degraded speech. Since these procedures lead to zero-phase filters, it is again only the spectral amplitude that is enhanced, with the phase of the filter being identical to that of the degraded speech.

### 2.2.1 Speech enhancement based on direct estimation of short-time spectral amplitude

A classical noise reduction approach for speech enhancement and robust recognition is the spectral subtraction method that was first proposed by Boll [1]. The basic idea is to restore the magnitude spectrum or power spectrum of a signal observed in additive noise through subtraction of an estimate of the average noise spectrum from the noisy signal spectrum. Assuming that the noise is a stationary or a slowly varying process, the noise spectrum is estimated or updated during the periods when the speech signal is absent. The estimation is performed on a frame-by-frame basis, where each frame consists of 20-40 ms of speech samples. The sample spectrum of the noisy signal is usually employed in the spectral subtraction approach, thus resulting in an estimate of the sample spectrum of the clean signal. An estimate of the sample autocorrelation function of the clean signal is obtained from the inverse discrete Fourier transform (IDFT) of the estimate of sample spectrum. The square root of the estimate of the sample spectrum is considered an estimate of the magnitude spectrum of the speech signal. An estimate of signal is obtained by combining the estimate of magnitude spectrum with the complex exponential of the phase of the noisy signal.

A block diagram of the spectral subtraction approach is shown in Fig. 2.1. The noisy signal  $x(n)$  is given by  $x(n) = s(n) + v(n)$ , where  $s(n)$  denotes the

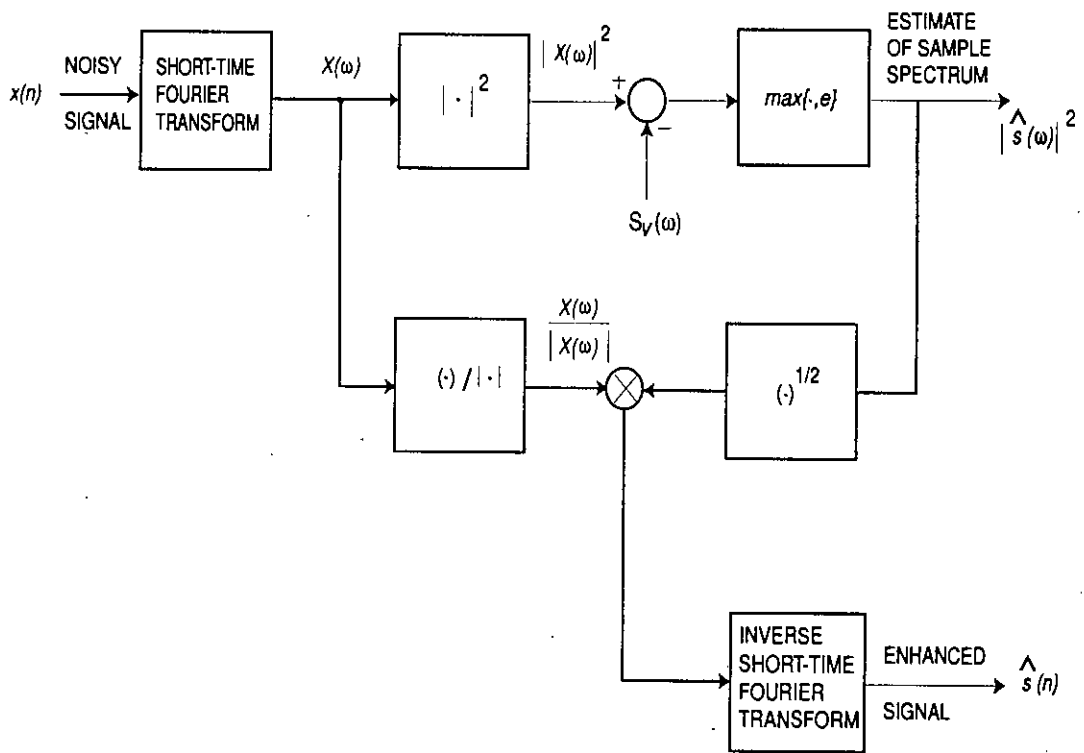


Fig. 2.1: The spectral subtraction approach

clean signal,  $v(n)$  denotes the additive noise sequence and it is assumed that  $s(n)$  and  $v(n)$  are uncorrelated. The short time Fourier transform (STFT) of the noisy signal  $x(n)$  [38], is denoted by  $X(\omega)$ ,  $0 \leq \omega \leq 2\pi$ . The sample spectrum of  $x(n)$  is given by  $|X(\omega)|^2$ . The estimate of the power spectral density of the noise process  $v(n)$  is denoted by  $S_v(\omega)$ . An estimate  $\hat{S}(\omega)$  of the Fourier transform of  $s(n)$  is then given by

$$\hat{S}(\omega) = [|X(\omega)|^2 - S_v(\omega)]^{1/2} \frac{X(\omega)}{|X(\omega)|} \quad (2.1)$$

provided that the difference of spectral estimates of the noisy signal and the noise process is nonnegative. If this difference becomes negative, then it is usually replaced by an arbitrary small nonnegative number, say  $\epsilon$ . The power spectral density of the noise is normally estimated from portions of the noisy signal during which speech is absent and only noise is present. The spectral subtraction based signal estimator affects the magnitude spectrum of the noisy signal in each frame while it keeps the phase of that signal intact. From a perceptual point of view this is a desirable property, since the short-time magnitude spectrum of the clean signal is considerably more important than its short-time phase spectrum [2], [3], [39], and optimal estimation of the short-time magnitude and phase spectrum of

the clean signal cannot be simultaneously performed [15], [25].

Many variations on the basic spectral subtraction approach have been proposed [1], [2], [18], [40]. The most popular modifications are those that involve averaging or smoothing of the estimate of the sample spectrum [1] controlling the amount of subtracted noise [2] and using different degrees of nonlinearity in estimating the magnitude spectrum of the clean signal [2], [40]. The latter two modifications are accomplished by the estimator reported in [2] as

$$\hat{S}(\omega) = [ |X(\omega)|^a - bE\{|V(\omega)|^a\} ]^{1/a} \frac{X(\omega)}{|X(\omega)|} \quad (2.2)$$

where  $V(\omega)$  is the Fourier transform of the realization of the noise process,  $a > 0$ , and  $b > 0$ . Eq. (2.2) degenerates to the standard spectral subtraction estimator given by Eq. (2.1) when  $a = 2$  and  $S_v(\omega) \approx E\{|V(\omega)|^2\}$ .

The spectral subtraction estimation approach has an intuitive basis and is relatively easy to implement. The advantage of the spectral subtraction method is its simplicity and effectiveness. The major calculation in the spectral subtraction method is the discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT) which can be efficiently implemented using the fast Fourier transform (FFT) algorithms. It is also proved by various researchers that the spectral subtraction method can improve the signal-to-noise ratio (SNR) and word recognition accuracy (WRA) under different SNR conditions [41]. The main problem in spectral subtraction is the processing distortions caused by the random variations of the noise spectrum and the use of noisy phase. For example, in order to avoid the negative estimates of the signal spectrum, the spectral subtraction output is usually post-processed by a mapping function  $T[\cdot]$  of the form

$$T[|\hat{S}(\omega)|] = \begin{cases} |\hat{S}(\omega)| & \text{if } |\hat{S}(\omega)| > g|X(\omega)|, \\ g|X(\omega)| & \text{otherwise} \end{cases} \quad (2.3)$$

where  $g$  is a positive parameter determined by the experiment. The distortion caused by the nonlinear mapping in Eq. (2.3) will produce a metallic sounding distortion, known as the "musical noise" which will be harmful to the automatic speech recognition (ASR) system especially when the input SNR is low.

An approach which leads to a further modification of spectral subtraction was proposed by McAulay and Malpass [19]. In this approach, the problem was formulated by assuming that at each frequency the noise is Gaussian and

developing the maximum likelihood estimate of  $|S(\omega)|$ . The resulting estimate has the form

$$|\hat{S}(\omega)| = \frac{1}{2}|X(\omega)| + \frac{1}{2}[|X(\omega)|^2 - E\{|V(\omega)|^2\}]^{1/2} \quad (2.4)$$

A further variation, proposed by McAulay and Malpass [19] modifies Eq. (2.4) by a factor which is chosen according to the probability of speech presence or absence.

### 2.2.2 Speech enhancement techniques based on Wiener filtering

In the previous section, the basis for enhancement was the explicit estimation of the short-time magnitude spectrum through a process of spectral subtraction. In this section, a frequency weighting for an "optimum" filter is first estimated from the noisy speech. This filter is then applied either in the time domain or frequency domain to obtain an estimate of the undegraded speech. As  $X(\omega)$ ,  $S(\omega)$  and  $V(\omega)$  denote the short-time spectra associated with the time functions  $x(n)$ ,  $s(n)$  and  $v(n)$ , the estimate  $\hat{S}(\omega)$  of  $S(\omega)$  takes the form

$$\hat{S}(\omega) = H(\omega)X(\omega) \quad (2.5)$$

As is well known, for  $x(n) = s(n) + v(n)$  in which  $s(n)$  and  $v(n)$  represent uncorrelated stationary random processes, the linear estimator of  $s(n)$  which minimizes the mean-square error is obtained by filtering  $x(n)$  with the noncausal Wiener filter. The noncausal Wiener filter cannot be applied directly to estimate  $s(n)$  since speech cannot be assumed to be stationary and the spectrum of the clean signal cannot be assumed known. An approach often used is to approximate the noncausal Wiener filter with an adaptive Wiener filter with frequency response

$$H(\omega) = \frac{E[|S(\omega)|^2]}{E[|S(\omega)|^2] + E[|V(\omega)|^2]} \quad (2.6)$$

The function  $E[|V(\omega)|^2]$  may be obtained either from the assumed known statistics of  $v(n)$  or by averaging many frames of  $|V(\omega)|^2$  during silence intervals in which the statistics of the background noise can be assumed to be stationary.  $E[|S(\omega)|^2]$  may be approximated as  $|\hat{S}(\omega)|^2$  or by smoothing  $|\hat{S}(\omega)|^2$  where  $|\hat{S}(\omega)|^2$  is obtained from the short-time spectral amplitude estimation scheme discussed in the previous section.

Given  $H(\omega)$ , the short-time speech segment is then obtained as specified by Eq. (2.5) applied either in the time domain or in the frequency domain. It should be noted that in all of the above procedures, the frequency weighting  $H(\omega)$  has zero phase and thus from Eq. (2.5) the phase associated with the estimate of  $\hat{S}(\omega)$  is that of  $X(\omega)$ .

Generalizations of Wiener filtering may also be considered. One such generalization which has been studied extensively in the context of image restoration has the frequency response given by [42]

$$H(\omega) = \left[ \frac{E[|\hat{S}(\omega)|^2]}{E[|S(\omega)|^2] + cE[|V(\omega)|^2]} \right]^d \quad (2.7)$$

for some constants “ $c$ ”, “ $d$ ” and has been referred to as parametric Wiener filters. By varying the constants “ $c$ ” and “ $d$ ”, filters with different characteristics can be obtained.

In the Wiener filter of Eq. (2.6) or its generalized form of Eq. (2.7) it is assumed that the term representing  $E[|S(\omega)|^2]$  is first obtained and the frequency weighting is then applied to  $X(\omega)$ , i.e.,

$$\hat{S}(\omega) = \left[ \frac{E[|S(\omega)|^2]}{E[|S(\omega)|^2] + cE[|V(\omega)|^2]} \right]^d X(\omega). \quad (2.8)$$

## 2.3 Speech Enhancement Techniques Based on Speech Model

A digital model of sampled speech that has been used in a number of practical applications and has a basis [43] in the physics of speech production system was shown in Fig. 2.2. In the model, the excitation source is either a quasi-periodic train of pulses for voiced sounds or random noise for unvoiced sounds. The digital filter represents the effects of the vocal tract, lip radiation and for voiced sounds, the glottal source. Since the vocal tract changes in shape as a function of time, the digital filter in Fig. 2.2 is in general time varying. However, over a short interval of time, the digital filter may be approximated as a linear time invariant system. Many systems which capitalize on the underlying speech model discussed in the preceding have been proposed in the literature for speech enhancement.

In the speech enhancement technique based on an underlying speech model, the parameters of the speech model are first estimated and then speech is gen-

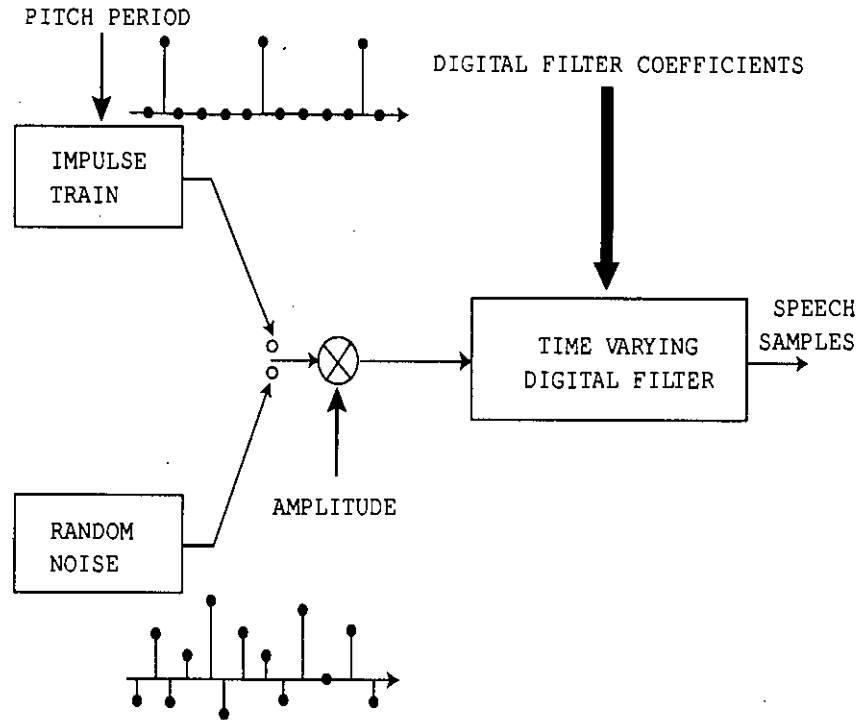


Fig. 2.2: A speech production model

erated based on the estimated parameters. The parameters of the model consist of the source parameters (pitch information) and the system parameters (vocal-tract information). Given the estimated parameters of a speech model, speech can be generated by a synthesis system based on the same underlying speech model or by designing a filter with the estimated speech model parameters and then filtering the noisy speech. The former approach requires both the source and system parameters while the latter approach generally requires only the system parameters. The techniques required for estimating the system parameters of a speech model, of course, depend on the specific model assumed. Even for the same speech model, however, there are again a variety of different techniques such as all-pole or pole-zero model of the vocal-tract, nonparametric speech models, etc. that may be used in estimating the model parameters.

The recent research on the model-based speech enhancement approaches use composite source models for the signal and noise. The composite source model is the most general statistical model known for speech signals, and it has proven extremely useful in speech recognition and enhancement applications. This model can also be useful for a wide class of noise sources encountered in practice, e.g.,



wideband noise, a mixture of noise sources, and competitive speech. A parametric model for the speech signal must be capable of providing a reasonable representation of at least the second-order statistics of this signal. These statistics have been shown to be extremely useful in speech processing as they can be used for synthesizing intelligible signals [44] and for recognizing speech signals [45]. By the second-order statistics we mean the different spectra of speech signals as well as the time-frequency correlation of those signals. This correlation can be extremely useful for speech enhancement applications, since it imposes smoothness constraints and thus significantly improves the robustness of the signal estimator.

## 2.4 Wavelet Speech Enhancement Based on the Teager Energy Operator

Wavelet transform has recently been evolved as a powerful tool for removing noise from speech and image signals. Bahoura and Rouat [37] have recently proposed a wavelet speech enhancement technique using the teager energy operator. The main idea was to define a discriminative threshold in various scales as a function of speech components.

In general, the measurements of a clean speech signal  $s(n)$  are corrupted by noise. Usually, the noise  $v(n)$  is modeled as an additive white Gaussian process with zero-mean and variance  $\sigma_v^2$ . The noisy speech signal  $x(n)$  is then given by

$$x(n) = s(n) + v(n), \quad n = 1, 2, \dots, N \quad (2.9)$$

### 2.4.1 Wavelet packet analysis

For a given level  $j$ , the wavelet packet ( $WP$ ) transform decomposes the noisy signal  $x(n)$  into  $2^j$  subbands corresponding to wavelet coefficients sets  $w_{k,m}^j$  as given by [46]

$$w_{k,m}^j = WP\{x(n), j\} \quad (2.10)$$

In other words,  $w_{k,m}^j$  represents the  $m$ th coefficient of the  $k$ th subband, where  $m = 1, 2, \dots, N/2^j$  and  $k = 1, 2, \dots, 2^j$ . For this application,  $WP$  decomposes the given signal at level 4 over which this method is applied.

### 2.4.2 Teager energy approximation

The Teager energy operator (TEO) is a powerful nonlinear operator proposed by Kaiser [47], capable of extracting the signal energy based on mechanical and physical considerations. For a bandlimited signal  $x(n)$ , this operator can be approximated by

$$\Omega_d[x(n)] = [x(n)]^2 - x(n+1)x(n-1). \quad (2.11)$$

The discrete-time teager energy operator (TEO) is applied to the resulting wavelet coefficients  $w_{k,m}^4$

$$t_{k,m}^4 = \Omega_d[w_{k,m}^4] \quad (2.12)$$

This operation enhances the discriminability of speech coefficients among those of noise.

### 2.4.3 Masks construction

For each subband coefficients, an initial mask is obtained by smoothing the TEO coefficients

$$M_{k,m}^4 = t_{k,m}^4 * h_k(m) \quad (2.13)$$

where  $h_k$  is a second order IIR lowpass filter.

### 2.4.4 Threshold modulation criterion

Ideally the standard threshold should be adapted only for speech frames and kept unchanged for nonspeech ones. The speech presence is interpreted by a significant contrast between peaks and valleys of  $M_k^4$ , while its absence is observed with a weaker contrast (smoother masks). To distinguish these frames, a parameter  $S_k^4$  named offset is defined, which estimates the valley's level. It is given by the abscissa of the maximum of the amplitude distribution  $H$  of the corresponding mask  $M_{k,m}^4$  and is estimated over the analyzed frame

$$S_k^4 = \text{abscissa}[\max(H(M_{k,m}^4))] \quad (2.14)$$

This parameter is close to 0 for speech frames and close to 1 for noise ones. If  $S_k^4$  is below the discriminatory value of  $0.35\max(M_{k,m}^4)$ , then the threshold is modulated or it remains unchanged.

### 2.4.5 Mask processing for the time-adapted threshold

The modulated threshold is then adapted to speech waveform by suppressing the offset and normalization before applying a root power function

$$M_{k,m}^{l4} = \left[ \frac{M_{k,m}^4 - S_k^4}{\max(M_{k,m}^4 - S_k^4)} \right]^{1/8} \quad (2.15)$$

### 2.4.6 Time-adapted threshold

Donoho and Johnstone [48] proposed a universal threshold  $\lambda$  for removing added white noise

$$\lambda = \sigma \sqrt{2 \log(N)} \quad (2.16)$$

with  $\sigma = MAD/0.6745$  where  $\sigma$  is the noise level. Median absolute deviation ( $MAD$ ) is estimated in the first scale. In the wavelet packet case, the threshold becomes

$$\lambda = \sigma \sqrt{2 \log(N \log_2 N)}. \quad (2.17)$$

For a given subband  $k$ , the time adapted threshold is defined as

$$\lambda_{k,m} = \lambda(1 - \xi M_{k,m}^{l4}) \quad (2.18)$$

where  $\xi$  is an adjustment parameter ( $\xi = 1$ ).

### 2.4.7 Thresholding $WP$ coefficients

The soft thresholding function is defined as [31], [46]

$$Ts(\lambda, w_k) = \begin{cases} \text{sign}(w_k)(|w_k| - \lambda), & \text{if } |w_k| \geq \lambda \\ 0, & \text{if } |w_k| < \lambda \end{cases} \quad (2.19)$$

where  $w_k$  represents the wavelet coefficients. The soft thresholding is then applied to the wavelet packet coefficients

$$\hat{w}_{k,m}^4 = Ts(\lambda_a, w_{k,m}^4) \quad (2.20)$$

where  $\lambda_a$  is the threshold corresponding to the analyzed frame

$$\lambda_a = \begin{cases} \lambda_{k,m}, & \text{if } S_k^4 \leq 0.35 \max(M_{k,m}^4) \\ \lambda, & \text{if } S_k^4 > 0.35 \max(M_{k,m}^4) \end{cases} \quad (2.21)$$

### 2.4.8 Inverse transformation

The enhanced signal is synthesized with the inverse transformation  $WP^{-1}$  of the modified  $WP$  coefficients

$$\hat{s}(n) = WP^{-1}(\hat{w}_{k,m}^A, j). \quad (2.22)$$

## 2.5 Conclusion

In this chapter some milestones in the development of speech enhancement have been reviewed. The general principle of speech enhancement based on the estimation of the short-time spectral amplitude of the speech is first discussed. This basic principle encompasses a variety of techniques and systems including the specific methods of spectral subtraction, parametric Wiener filtering, etc. Then a variety of systems that rely on more specific modeling of the speech waveform have been discussed briefly. Lately, discrete cosine transform (DCT) and wavelet transform have been widely used as analysis tools in the field of speech enhancement. An wavelet transform based enhancement method proposed by Bahoura and Rouat [37] is covered at the end of this chapter. In subsequent chapters a new method is proposed as an improved denoising technique which is equally applicable to both wavelet and DCT domain and the performance of this method will also be reported with necessary evaluations.

# Chapter 3

## Bias-compensated Noise Level for Wavelet and DCT Speech Enhancement

### 3.1 Introduction

Speech samples are usually corrupted by noise in the real world. To reduce the influence of noise, two research topics - the speech enhancement and speech recognition in noisy environments - have arose. For speech enhancement, i.e., the extraction of a signal corrupted by noise, several approaches are available. Among them, spectral subtraction [1]-[20] is one of the most popular single-channel speech enhancement methods due to its computational efficiency. Despite its capability of removing background noise, this method introduces musical noise. Speech enhancement has also been accomplished by modifying the temporal contours of the parameters or features, like spectral band energies [22]. In recent years several alternative approaches such as signal subspace methods [23]-[24] and HMM-based algorithms [6], [25]-[26] have been proposed for enhancing degraded speech. Methods for speech enhancement have also been developed based on extraction of parameters from noisy speech, and synthesizing speech from these parameters [27], [28]. Recently, discrete cosine transform (DCT) and wavelet transform have been widely used as analysis tools in the field of speech enhancement [30]-[36].

Bahoura and Rouat [37] have recently proposed a speech enhancement technique in the wavelet domain. The main idea was to define a time adapted threshold in various scales as a function of speech components. Setting of a threshold criterion requires an accurate estimate/knowledge of the additive noise level in

the noisy speech. In speech, most of the signal energy is usually retained in the lower frequency range. The higher frequency region of the transform coefficients mostly contains noise from which noise level is usually estimated [48]. The effect of the signal components present in this region may be insignificant at low SNRs but is not negligible particularly at high SNRs. For this reason, this method shows deteriorating performance at a relatively high SNR. As for example, a signal having an SNR of 20 dB deteriorated to an SNR of 16.47 as shown in Table I of [37].

In this research, we propose an improved speech enhancement method in both wavelet and DCT domain. Unlike other conventional techniques, more accurate estimate of the threshold parameter, the noise level, is obtained by compensating the effect of the trace of the signal remaining at the high frequency region of the transform coefficients of the degraded speech. Using fourth-order statistics, we introduce a novel approach for computing a correction factor to reduce the upward bias of the noise level obtained from the median absolute deviation (*MAD*) of the high frequency coefficients of the degraded speech.

## 3.2 Problem Formulation

In general, the measurements of a clean speech signal  $s(n)$  are corrupted by noise. Usually, the noise  $v(n)$  is modeled as an additive white Gaussian process with zero-mean and variance  $\sigma_v^2$ . The noisy speech signal  $x(n)$  is then given by

$$x(n) = s(n) + v(n), \quad n = 1, 2, \dots, N \quad (3.1)$$

The objective of this research is to extract the speech signal  $s(n)$  from the degraded observed signal  $x(n)$  by applying a new speech enhancement technique based on the combined application of hard and soft thresholding of transform coefficients of the noisy signal with bias-compensated noise level as the threshold parameter. We investigate the performance of the proposed method in both wavelet and DCT domain.

### 3.3 Estimation of Noise Level: Conventional Approach

According to Donoho and Johnstone [48], the noise level is defined by

$$\sigma_{v^+} = MAD/0.6745 \quad (3.2)$$

where  $MAD$  (median absolute deviation) is computed from the wavelet coefficients at the finest level. In DCT based analysis,  $MAD$  of the DCT coefficients is computed at the finest level ( $N/2 + 1$  to  $N$ ) to estimate the threshold parameter for the denoising process. Usually, the coefficients at the finest level are predominantly noise. Because of the presence of a small fraction of signal coefficients, we get an estimate of the noise level that suffers an upward bias [48]. To signify this fact the subscript “ $v^+$ ” is used instead of just “ $v$ ” in Eq. (3.2).

Fig. 3.1. (a) and (b) shows the variation of  $\sigma_{v^+}$  with SNR for a given noisy sequence in the wavelet and DCT domain, respectively. The curve is obtained by adding white noise to a given speech signal at various SNR levels, and that the corresponding  $\sigma_{v^+}$  is calculated using Eq. (3.2). It is interesting to observe that  $\sigma_{v^+}$  shows asymptotically flat behaviour for higher values of SNR. This indicates that the coefficients at the finest level contains signal whose  $MAD$  corresponds to this asymptotic value. Applying  $\sigma_{v^+}$  as a threshold parameter removes some of the signal components which have significant adverse effect at high SNRs. For this reason, most of the existing methods concerning denoising of the speech signal encounter a strong drawback of SNR reduction of the denoised speech [37]. This is due to undesired inclusion of the effect of the signal components while calculating  $\sigma_{v^+}$ . As for example, as shown in Table I of [37] the SNR of the enhanced speech is found to be 14.47 and 16.47 dB while the original SNR of the noisy speech was 15 and 20 dB, respectively. This justifies our observation on the behaviour of  $\sigma_{v^+}$  depicted in Fig. 3.1 and suggests that introduction of a correction factor in  $\sigma_{v^+}$  is necessary to make the threshold value more effective.

### 3.4 Wavelet Transform Based Proposed Enhancement Algorithm

For a given level  $j$ , the wavelet packet ( $WP$ ) transform decomposes the noisy signal  $x(n)$  into  $2^j$  subbands corresponding to wavelet coefficients sets  $X_{k,m}^j$  as

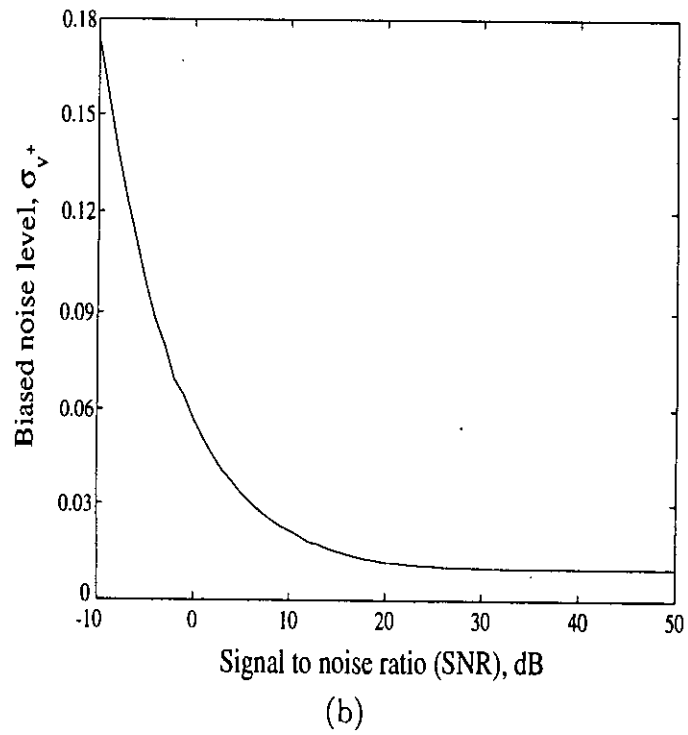
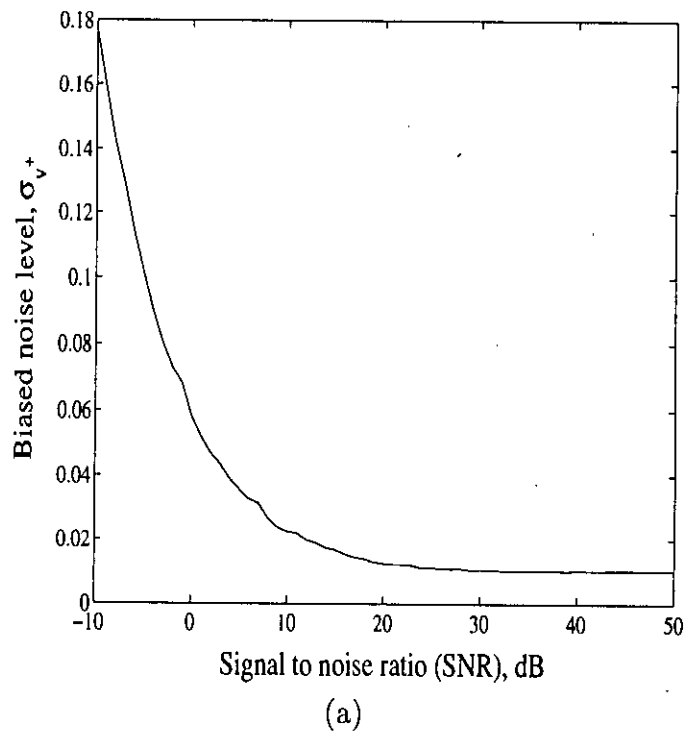


Fig. 3.1: Variation of  $\sigma_{v+}$  with SNR: (a) wavelet; (b) DCT.



given by [46]

$$W_{k,m}^j = WP\{x(n), j\} \quad (3.3)$$

In other words,  $w_{k,m}^j$  represents the  $m$ th coefficient of the  $k$ th subband, where  $m = 1, 2, \dots, N/2^j$  and  $k = 1, 2, \dots, 2^j$ . For this application,  $WP$  decomposes the given signal at level 4 over which the proposed method is applied. But in estimating the noise level,  $WP$  transform of the degraded speech at first scale is used.

### 3.4.1 Calculation of corrected value of noise level

In this research, the noise level used as the threshold parameter is estimated taking into account the speech component present at the finest level. It is known that for white Gaussian noise the kurtosis is 3 [49]. On the other hand, the distribution of signal coefficients remaining at the finest level is sharply peaked, i.e., leptokurtically distributed with kurtosis much larger than 3. Thus, at the finest region the kurtosis gradually decreases with increasing noise to a given speech and asymptotically reaches 3 when noise is much greater than signal. Therefore, kurtosis can be used to estimate the correction factor in proportional to the signal present in the  $MAD$  of the wavelet coefficients at the finest level. The kurtosis of wavelet coefficients is computed as [49]

$$K_4 = \frac{E\{[W_{k,m}]^4\}}{(E\{[W_{k,m}]^2\})^2} \quad (3.4)$$

where  $E[\cdot]$  denotes the expectation operator. The normalized kurtosis defined in [21] is given by

$$\gamma_4 = \frac{E\{[W_{k,m}]^4\}}{(E\{[W_{k,m}]^2\})^2} - 3 \quad (3.5)$$

It may be mentioned that the lower-bound on  $\gamma_4$  is zero for wavelet coefficients of a purely white Gaussian noise sequence, the upper limit of  $\gamma_4$  is a function of the signal present at the finest level and may vary widely from speech to speech, and lies in between for different noisy versions with varying noise level of a given speech. Therefore, scaling of  $\gamma_4$  is necessary to re-normalize it between 0 and 1 for a given speech under different noisy conditions. Here, we deliberately add different white Gaussian noise sequences of increasing strength to the given noisy speech signal. Let  $\gamma_4(\zeta_p)$  denotes the normalized kurtosis after the addition of the  $p$ th auxiliary noise sequence with the given noisy speech so that the SNR (in dB)

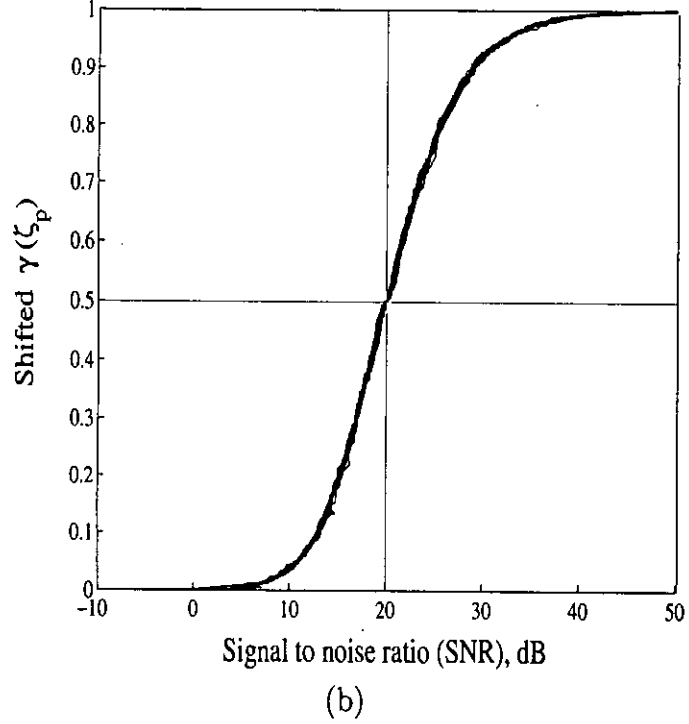
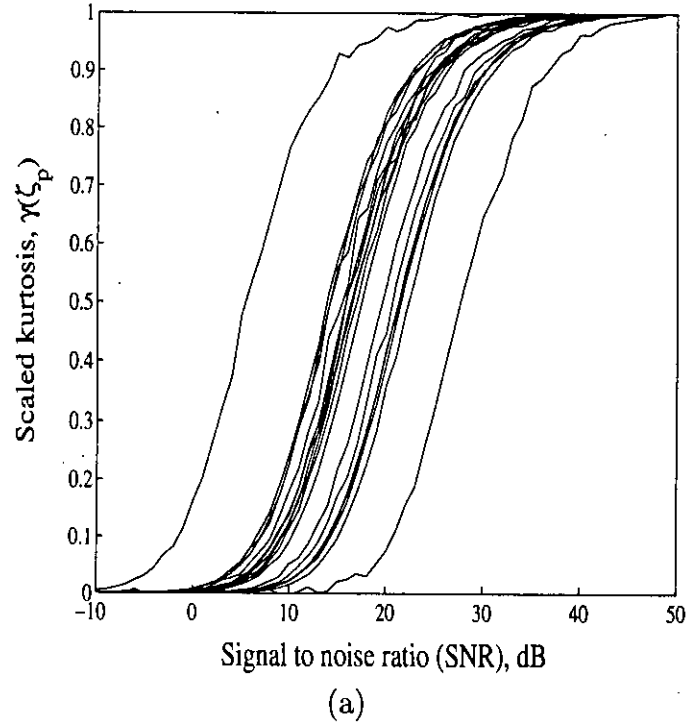


Fig. 3.2: Variation of scaled kurtosis with SNR in wavelet domain: (a)  $\gamma(\zeta_p)$  vs. SNR; (b) Shifted  $\gamma(\zeta_p)$  vs. SNR.

is  $\zeta_p$  with  $\zeta_0$  for  $p = 0$  being the SNR of the original given noisy speech. Then, the scaled values of kurtosis,  $\gamma(\zeta_p)$ , may be obtained as

$$\gamma(\zeta_p) = \frac{\gamma_4(\zeta_p)}{\gamma_4(\zeta_0)}, \quad p = 0, 1, 2, \dots, P \quad (3.6)$$

We select arbitrarily 20 clean speeches from the TIMIT and other standard databases to investigate the behaviour  $\gamma(\zeta_p)$  under different noisy conditions. The variation of  $\gamma(\zeta_p)$  with SNR is shown in Fig. 3.2 (a). It is interesting to note that all curves show asymptotically flat behaviour giving 1 and 0 for higher and lower values of SNR, respectively. This complies with the presumption that at a high SNR, the high frequency region of the noisy signal contains predominantly speech coefficients with negligible noise coefficients, whereas signal coefficients can be considered negligible in comparison to that of noise at low SNR values.

We shift each curve in Fig 3.2 (a) horizontally such that  $\gamma(\zeta_p) = 0.5$  occurs at an SNR  $\zeta_p = 20$  dB for each curve without loss of generality. The results are shown in Fig. 3.2 (b). It is found that this transformation gives a single curve with negligible thickness. Therefore, different  $\gamma(\zeta_p)$  curves in Fig. 3.2 (a) are actually identical in shape. This shifting is only due to the variation of the amount of speech component present at the finest level for different speeches. Now, an empirical function that best fits the curve in Fig. 3.2 (b) can be obtained as

$$\Gamma(\zeta_p) = \frac{1}{1 + e^{-\alpha(\zeta_p - 20)}} \quad (3.7)$$

where  $\alpha$  is numerically found to be 0.27. Then any curve in Fig. 3.2(a) may be obtained simply by shifting  $\Gamma(\zeta_p)$  as  $\Gamma(\zeta_p - i)$  where  $i$  is the shift parameter in SNR. This function  $\Gamma(\zeta_p)$  would be termed as the template function subsequently.

The template function defined in Eq. (3.7) is used as a basis function for determining the correction factor in order to reduce the undesired signal-bias included in  $\sigma_{v+}$ . This is done in the following way. First, a curve similar to the template function is generated from the given noisy signal. In a real situation, only a single value  $\gamma_4(\zeta_0)$  of the given noisy signal can be calculated. However, a set of  $\gamma_4(\zeta_p)$  can also be found by adding computer generated known white noise sequences with increasing power to the given noisy sequence. The process of adding auxiliary noise sequences with the given signal is terminated when  $\gamma_4(\zeta_p)$  reaches zero (i.e.,  $\gamma_4(\zeta_p) \leq 0$ ). Scaling these values as in Eq. (3.6) generates a

curve which is denoted here as  $\gamma(\zeta_p)$ . Second, the degree of similarity between the template function  $\Gamma(\zeta_p)$  and the generated curve  $\gamma(\zeta_p)$  is measured for determining the signal proportion in the mixture of signal plus noise coefficients at the finest level. The curve  $\gamma(\zeta_p)$  generated in the preceding way closely resembles the template function  $\Gamma(\zeta_p)$ , though may be shifted in SNR, if the SNR of the given noisy speech ( $\zeta_0$ ) is very high. Because, the whole curve as that of the template function can be generated adding auxiliary noise to the given noisy speech except a possible shift in SNR due to difference in the speech signals. On the other hand, for a very low SNR the degree of mismatch is very high as  $\gamma(\zeta_p)$  resembles only a small portion of the bottom part of the template function. For an intermediate SNR of the given noisy signal, the degree of similarity between the curves  $\gamma(\zeta_p)$  and  $\Gamma(\zeta_p)$  vary in between the above two cases. In this research, the maximum value of the cross-correlation between the curves  $\gamma(\zeta_p)$  and  $\Gamma(\zeta_p)$  is used as a measure of the degree of similarity. Because, irrespective of the shape,  $\gamma(\zeta_p)$  is in general shifted from  $\Gamma(\zeta_p)$  in SNR and thus the maximum value of cross-correlation is achieved when they overlap. The cross-correlation between the two functions  $\gamma(\zeta_p)$  and  $\Gamma(\zeta_p)$  is defined [50] as

$$R_{\gamma\Gamma}(d) = \frac{1}{P} \sum_{p=1}^P \Gamma(p)\gamma(p-d), \quad d = 0, 1, 2, \dots \quad (3.8)$$

where  $\gamma(p)$  refers to  $\gamma(\zeta_p)$  and  $\Gamma(p)$  refers to  $\Gamma(\zeta_p)$  at the  $p$ th instant of auxiliary noise addition and  $P$  denotes the number of points used to generate the template function. The maximum value of the function  $R_{\gamma\Gamma}(d)$ , denoted by  $R_{max}$ , indicates the degree of similarity between the template and the test functions.

The values of  $R_{max}$  are examined using different known speeches for various SNRs. It is observed that for very high values of SNR, the value of  $R_{max}$  is  $\approx 0.48$  and is almost independent of speech signal. Let us denote this value by  $\Psi$ . On the other hand, when the SNR of the given noisy speech is very low,  $R_{max}$  is found to be  $\leq 0.255$ . However, under such a low SNR value the effect of signal bias is insignificant on the estimated value of the noise level. Therefore, we choose 0.255 to be a practical lower limit for  $R_{max}$  at a very low SNR. This is denoted as  $\psi$ . When  $R_{max}$  is  $\Psi$ ,  $\sigma_v^+$  in Eq. (3.2) solely gives the signal level. Hence 100% compensation is needed. On the contrary, when  $R_{max}$  is  $\psi$ ,  $\sigma_v^+$  is completely dominated by the noise level and no compensation is required. If the value of

$R_{max}$  lies between  $\Psi$  and  $\psi$ , a linear interpolation for the proposed correction factor denoted by  $\beta$  is assumed and is defined by

$$\beta = \begin{cases} \min\left(1, \frac{R_{max} - \psi}{\Psi - \psi}\right), & \text{if } R_{max} \geq \psi \\ 0, & \text{otherwise} \end{cases} \quad (3.9)$$

where  $\beta$  is intended to limit in the interval  $[0, 1]$ . The corrected noise level  $\sigma_v$  is now defined by

$$\sigma_v = (1 - \beta)\sigma_{v+}, \quad 0 \leq \beta \leq 1 \quad (3.10)$$

It is obvious that  $\sigma_v = 0$  when  $\beta = 1$  (i.e., high SNR case), and  $\sigma_v = \sigma_{v+}$  when  $\beta = 0$  (i.e., low SNR case), and compensate for the signal-bias in proportion to the signal component present at the finest level of the wavelet coefficients in between these two limiting cases as expected.

### 3.4.2 Thresholding $WP$ coefficients

In this research, conventional amplitude subtraction based soft thresholding alone and the successive application of hard and soft thresholding are applied on the  $WP$  coefficients at level 4. Details of the thresholding techniques are given in the following.

#### A. Application of soft thresholding alone

In line of the reported literature, soft thresholding is applied for enhancement of the noisy speech signal. The amplitude subtraction based soft thresholding technique is defined as [31], [46]

$$\tilde{W}_{k,m}^j = \begin{cases} \text{sign}(W_{k,m}^j)(|W_{k,m}^j| - \sigma_v), & \text{if } |W_{k,m}^j| \geq \sigma_v \\ 0, & \text{if } |W_{k,m}^j| < \sigma_v \end{cases} \quad (3.11)$$

Notice that unlike other methods [31], [46], the corrected noise level  $\sigma_v$  is used as the threshold parameter.

#### B. Combined application of hard and soft thresholding

We simultaneously apply both hard and soft thresholding to devise an improved denoising method. Regions in wavelet domain, where average signal strength is less than that of noise, hard thresholding is applied by forcing the coefficients to be zero. After accomplishing hard thresholding, soft thresholding is applied over the rest of the regions to further reduce the noise level.

The  $WP$  coefficients at a particular level is first divided into a number of blocks consisting of convenient number of consecutive  $WP$  coefficients. Then

hard thresholding is applied to a block of  $WP$  coefficients where average signal power is less than the average noise power as that block essentially contributes more noise than signal in the denoised speech. To identify the blocks for hard thresholding, a window of length same as the block size is slid over the whole range. The hard thresholding used in this paper is defined as

$$\bar{W}_{k,w}^j = \begin{cases} W_{k,w}^j, & \text{if } P_x^w \geq 2P_v^w \\ 0, & \text{if } P_x^w < 2P_v^w \end{cases} \quad (3.12)$$

where  $w = m$  to  $m + l - 1$ ,  $l$  is the length of the window and  $P_x^w (= P_s^w + P_v^w)$  represents the total power of the  $WP$  coefficients inside a given window and  $P_v^w$  denotes the power of the noise component over the same window. An estimated value of  $P_v^w$  can be obtained using the relation  $P_v^w = l\sigma_v^2$  and  $P_x^w$  can be estimated simply by taking the sum of the squared value of the  $WP$  coefficients for that given window.

The hard thresholding described earlier eliminates a significant portion of noise from the regions of wavelet coefficients where noise dominates signal. The rest of the regions where signal strength is higher than that of noise, soft thresholding is applied for further enhancement of the noisy signal. As the noise power uniformly penetrates into the actual signal in wavelet domain, subtraction of noise power from the signal power is expected to improve the SNR of the enhanced signal. The coefficients with power less than average noise power are more susceptible to distortion; their amplitudes are reduced proportionately and the soft thresholding applied in this paper is defined as

$$\tilde{W}_{k,m}^j = \begin{cases} \text{sign}(\bar{W}_{k,m}^j) \sqrt{|\bar{W}_{k,m}^j|^2 - \sigma_v^2}, & \text{if } |\bar{W}_{k,m}^j| \geq \sigma_v \\ \frac{\bar{W}_{k,m}^j |\bar{W}_{k,m}^j|}{\sigma_v}, & \text{if } |\bar{W}_{k,m}^j| < \sigma_v \end{cases} \quad (3.13)$$

where  $\sigma_v$  is the corrected noise level.

### 3.4.3 Reconstruction of the original signal

The enhanced signal is synthesized with the inverse transformation  $WP^{-1}$  of the modified  $WP$  coefficients ( $\tilde{W}_{k,m}^j$ ), i.e.,

$$\hat{s}(n) = WP^{-1}(\tilde{W}_{k,m}^j, j). \quad (3.14)$$

### 3.5 DCT Based Proposed Enhancement Algorithm

The topic on speech enhancement is widely researched and many speech enhancement algorithms make use of the Discrete Fourier Transform (DFT) to make it easier to remove noise embedded in the noisy speech signal. This is often done as it is easier to separate the speech energy and the noise energy in the transform domain. For example, the energy of white noise is uniformly spread throughout the entire spectrum, but the energy of speech, especially voiced speech, is concentrated in certain frequencies. Most of the algorithm only attempt to modify the spectral amplitudes of the noise corrupted speech signal in order to reduce the effect of noise component while leaving the noise corrupted phase information intact. It is of interest to note that in [15], the best estimate of the phase of the speech component was shown to be the phase of the corrupted signal itself. Hence, the advantage of using a real transform, such as Discrete Cosine Transform (DCT), is that the problem of not correcting for the phase will result in less severe consequences [30]. Also, the DCT has the added advantage of higher spectral resolution than the DFT for the same window size. For a window size of  $N$ , the DCT has  $N$  independent spectral components while the DFT only produces  $N/2 + 1$  independent spectral components, as the other components are just complex conjugates. Thus, the DCT outperforms the DFT [30].

The forward DCT of the noisy signal  $\{x(n), 0 \leq n \leq N - 1\}$  is given by [30]

$$X(k) = \alpha(k) \sum_{n=0}^{N-1} x(n) \cos \left[ \frac{\pi(2n+1)k}{2N} \right], \quad 0 \leq k \leq N-1 \quad (3.15)$$

where

$$\alpha(k) = \begin{cases} \sqrt{\frac{1}{N}}, & k = 0 \\ \sqrt{\frac{2}{N}}, & 1 \leq k \leq N-1 \end{cases} \quad (3.16)$$

We have to denoise the speech signal with enhanced SNR and better subjective performance by modifying the noisy DCT coefficients  $X(k)$  using a new technique with signal-bias compensated noise level as the threshold parameter. The reconstructed signal,  $\hat{s}(n)$ , can be obtained using the following inverse discrete

cosine transformation [30]

$$\hat{s}(n) = \sum_{k=0}^{N-1} \alpha(k) \tilde{X}(k) \cos \left[ \frac{\pi(2n+1)k}{2N} \right], \quad 0 \leq n \leq N-1 \quad (3.17)$$

where  $\tilde{X}(k)$  denotes the denoised DCT coefficients.

### 3.5.1 Calculation of corrected value of noise level

A similar procedure described in section 3.4.1 is adopted to estimate the correction factor  $\beta$  in the DCT domain. The corresponding scaled and shifted kurtosis curves obtained using DCT instead of wavelet transform is presented in Fig. 3.3. The biased noise level is then corrected according to the Eq. (3.10).

### 3.5.2 Thresholding DCT coefficients

As described in section 3.4.2, here we also apply two types of thresholding, namely, soft thresholding alone, and the combined application of hard and soft thresholding.

#### A. Application of soft thresholding alone

The amplitude subtraction based soft thresholding technique is defined as

$$\tilde{X}_k = \begin{cases} \text{sign}(X_k)(|X_k| - \sigma_v), & \text{if } |X_k| \geq \sigma_v \\ 0, & \text{if } |X_k| < \sigma_v \end{cases} \quad (3.18)$$

Notice that unlike other methods [31], [46], the corrected noise level  $\sigma_v$  is used as the threshold parameter.

#### B. Combined application of hard and soft thresholding

We simultaneously apply both hard and soft thresholding in the DCT domain using the bias-compensated noise level as the threshold parameter. The hard thresholding used here is defined as

$$\bar{X}(k) = \begin{cases} X(k), & \text{if } P_x^w \geq 2P_v^w \\ 0, & \text{if } P_x^w < 2P_v^w \end{cases} \quad (3.19)$$

where  $P_x^w$  represents the total power of the DCT coefficients inside a given window and  $P_v^w$  denotes the power of the noise component over the same window.

Hard thresholding eliminates a significant portion of noise from the regions of DCT coefficients where noise dominates signal. The rest of the regions where signal strength is higher than that of noise, soft thresholding is applied for further enhancement of the noisy signal. The soft thresholding is defined as



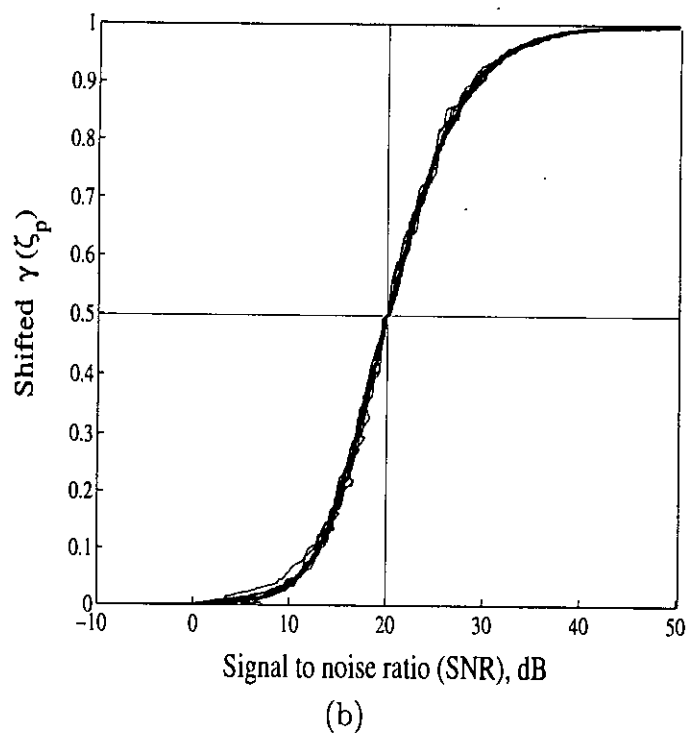
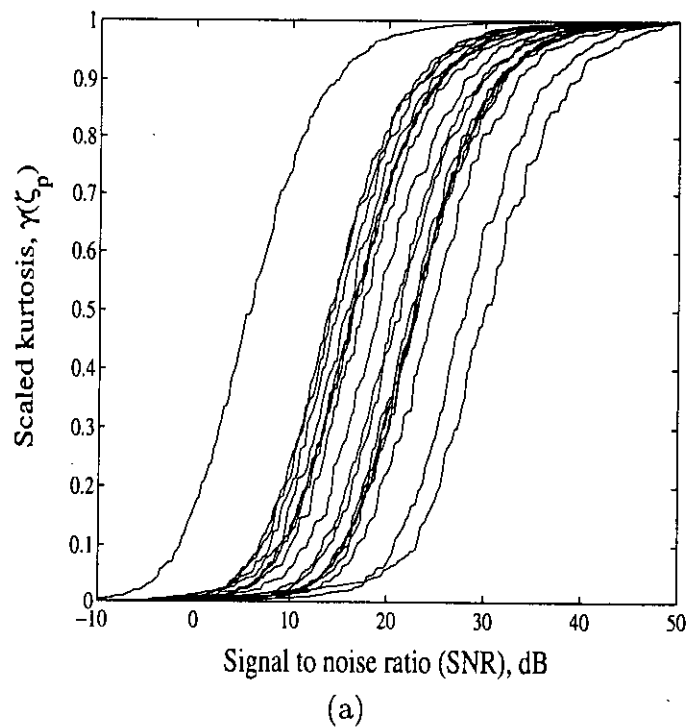


Fig. 3.3: Variation of scaled kurtosis with SNR in DCT domain: (a)  $\gamma(\zeta_p)$  vs. SNR; (b) Shifted  $\gamma(\zeta_p)$  vs. SNR.

$$\tilde{X}(k) = \begin{cases} \text{sign}(\bar{X}(k))\sqrt{|\bar{X}(k)|^2 - \sigma_v^2}, & \text{if } |\bar{X}(k)| \geq \sigma_v \\ \frac{\bar{X}(k)|\bar{X}(k)|}{\sigma_v}, & \text{if } |\bar{X}(k)| < \sigma_v \end{cases} \quad (3.20)$$

where  $\sigma_v$  is the corrected noise level.

### 3.6 Conclusion

Speech enhancement using a new thresholding technique in both wavelet and DCT domain has been proposed. The major focus of this chapter was to develop a method for better estimation of the noise level (i.e., signal-bias compensated noise level) considering the signal remaining at the finest level. The signal-bias is compensated by exploiting the behaviour of fourth-order statistics of the coefficients at the high frequency region. A new thresholding technique is then proposed that employs both hard and soft thresholding successively. Using corrected noise level as the threshold parameter, conventional amplitude subtraction based soft thresholding alone and the proposed thresholding technique in both wavelet and DCT domain have been applied. The performance of the thresholding techniques will be discussed with necessary measures in the following chapter.

# Chapter 4

## Results

### 4.1 Data Used

The proposed enhancement algorithm is tested for a data set consisting of 20 different continuous speech sentences from the TIMIT and other sources. Half of the sentences are spoken by female speakers while the remaining sentences are by male speakers. Simulations for four different clean speeches from the TIMIT database are reported for comparing the proposed method with the one described in [37]. The speech signals are sampled at 8 kHz and quantized to 16 bits. Noise types considered in our experiments include white Gaussian noise and real noise recorded inside a moving car with air cooler turned on. White noise was used for the following reasons: First, white noise affects the entire frequency band of the speech signals and is therefore considered one of the most perceptually harmful noise sources. Second, white noise is a good model for wideband noise sources which are often encountered in practice, e.g., thermal noise in communication systems. Third, white noise has been commonly used in studying the performance of enhancement systems and can therefore be seen as a “standard” test noise source.

### 4.2 Estimation of Corrected Noise Level

The SNR of the given noisy signal is to be estimated prior to the estimation of the correction factor ( $\beta$ ). The noisy speech is obtained by adding computer generated white Gaussian noise with the clean speech. For this, a female clean speech “She had your dark suit in greasy wash water all year” termed as s1 and a male speech “Should we chase those cowboys?” termed as s2 are taken from

Table 4.1: Comparison of actual and corrected noise levels along with the correction factor,  $\beta$ , for the speech, "She had your dark suit in greasy wash water all year", at different SNRs.

Method	SNR of given noisy speech (dB)	Estimated correction factor $\beta$	Biased noise level $\sigma_{v+}$	Corrected noise level $\sigma_v$	Actual noise level $\sigma_v^T$
Wavelet	-10	0.00	0.2973	0.2973	0.2993
	-5	0.00	0.1713	0.1713	0.1683
	0	0.07	0.0987	0.0916	0.0946
	5	0.09	0.0580	0.0528	0.0532
	10	0.17	0.0360	0.0299	0.0301
	15	0.28	0.0227	0.0164	0.0168
	20	0.42	0.0159	0.0092	0.0095
	25	0.54	0.0120	0.0055	0.0053
	30	0.62	0.0100	0.0038	0.0030
DCT	-10	0.00	0.3041	0.3041	0.2993
	-5	0.00	0.1690	0.1690	0.1683
	0	0.02	0.0982	0.0965	0.0946
	5	0.06	0.0580	0.0548	0.0532
	10	0.18	0.0365	0.0299	0.0301
	15	0.26	0.0239	0.0178	0.0168
	20	0.42	0.0175	0.0101	0.0095
	25	0.54	0.0129	0.0059	0.0053
	30	0.64	0.0112	0.0040	0.0030

the TIMIT database and then added with the computer generated white noise sequences. The noise power is first estimated using Eq. (3.2) and then the signal power is obtained by subtracting it from the noisy speech power. The values of the noise and signal power thus obtained give an under estimate of SNR due to the upward bias in the estimated noise level. However, this error is found to have insignificant effect on  $\beta$ . As shown in Fig. 4.1, the final estimate of SNR using the corrected noise level is much more accurate than the one using the biased noise level.

As described in section 3.4.1, we add computer generated auxiliary white Gaussian noise sequences of increasing power with the given noisy signal of 22220 (s1) and 15616 (s2) samples. For convenience, we choose the noise power such that it results in a decremental SNR of 1 dB. Note that samples from the template function has to be taken at the same interval. The addition of noise sequences with the given signal is terminated when  $\gamma_4(\zeta_p) \leq 0$  is satisfied. The SNR was found to be  $-5.51$  and  $-3.27$  dB at the termination condition for s1 and s2, respectively. Estimated results of  $\beta$  (Eq. (3.9)) at the high frequency region for different SNRs are presented in Tables 4.1 and 4.2 for two different utterances s1 and s2 along with the corrected noise level ( $\sigma_v$ ) using Eq. (3.10) and the actual noise level ( $\sigma_v^T$ ). It is seen that the corrected values are fairly close to the actual ones. Since  $\sigma_v$  determines the threshold level for the noisy coefficients, an over-estimation of  $\sigma_v^T$  would have an adverse effect on the denoised speech. In particular, Tables 4.1 and 4.2 show that the biased noise level  $\sigma_{v+}$  (Eq. (3.2)) is significantly higher than the proposed corrected noise level  $\sigma_v$  (Eq. (3.10)) at a relatively high SNR.

### 4.3 Performance Test

To evaluate the performance of the proposed algorithm, both objective and subjective tests are performed in wavelet and DCT domain. The non-stationarity of speech signals require that the duration of an analysis segment be of approximately  $20 \sim 40$  ms. But, since wavelet transform itself is a time-scale representation, it requires no segmentation of the noisy  $WP$  coefficients. For this reason, no segmentation is adopted during wavelet based analysis. During DCT based analysis, the given noisy speech is segmented so that each segment be of 32 ms

Table 4.2: Comparison of actual and corrected noise levels along with the correction factor,  $\beta$ , for the speech, "Should we chase those cowboys?", at different SNRs.

Method	SNR of given noisy speech (dB)	Estimated correction factor $\beta$	Biased noise level $\sigma_{v+}$	Corrected noise level $\sigma_v$	Actual noise level $\sigma_v^T$
Wavelet	-10	0.00	0.4066	0.4066	0.4045
	-5	0.00	0.2310	0.2310	0.2274
	0	0.00	0.1312	0.1312	0.1279
	5	0.09	0.0769	0.0698	0.0719
	10	0.13	0.0455	0.0396	0.0404
	15	0.25	0.0290	0.0219	0.0227
	20	0.36	0.0188	0.0121	0.0158
	25	0.53	0.0126	0.0059	0.0072
	30	0.68	0.0093	0.0030	0.0040
DCT	-10	0.00	0.3996	0.3996	0.4045
	-5	0.00	0.2283	0.2283	0.2274
	0	0.00	0.1311	0.1311	0.1279
	5	0.03	0.0763	0.0739	0.0719
	10	0.11	0.0475	0.0426	0.0404
	15	0.22	0.0306	0.0239	0.0227
	20	0.32	0.0208	0.0141	0.0158
	25	0.46	0.0166	0.0089	0.0072
	30	0.52	0.0105	0.0051	0.0040

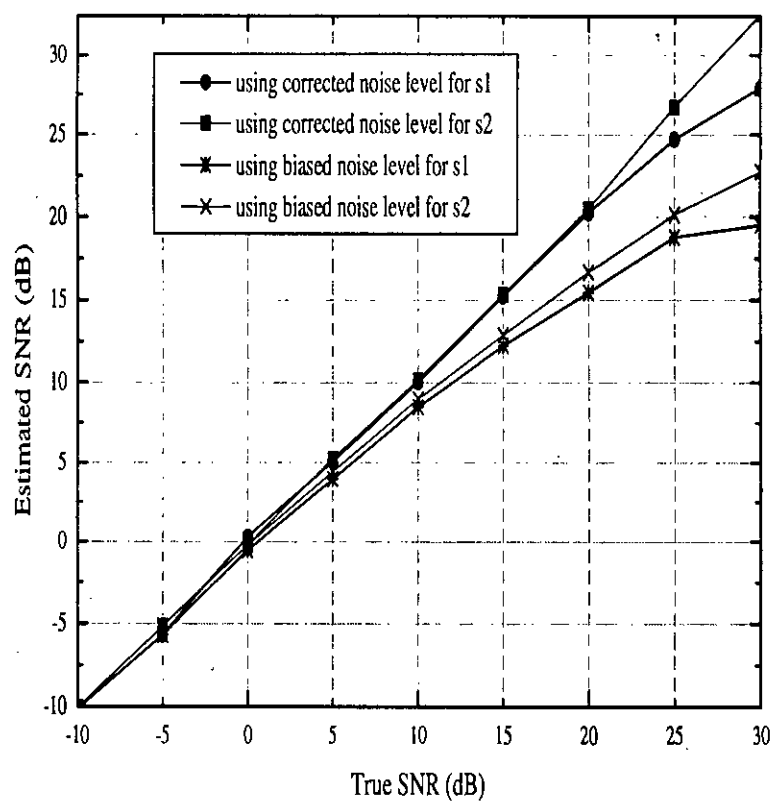


Fig. 4.1: Estimation of SNR of noisy speech.

Table 4.3: Results on SNR improvement for the speech, “Should we chase those cowboys?”, corrupted by additive white noise

SNR (dB)	Wavelet packet using Ref. [37] (dB)	Proposed wavelet packet method		Proposed DCT method	
		soft thr. alone (dB)	hard and soft thr. (dB)	soft thr. alone (dB)	hard and soft thr. (dB)
-10	-1.95	-1.25	1.78	-1.05	2.53
-5	1.84	1.53	4.66	3.17	5.64
0	5.92	5.56	7.81	6.85	9.10
5	9.68	9.50	11.20	10.76	12.49
10	13.09	13.53	14.70	11.67	16.11
15	15.68	17.54	18.55	18.63	19.71
20	18.01	21.69	22.47	22.52	23.43
25	20.13	26.04	26.54	26.70	27.23
30	22.14	30.57	30.75	31.07	31.01

duration, i.e. 256 samples, and the overlapping between two consecutive segments is taken to be 115 samples ( $\cong 45\%$  of segment duration).

### 4.3.1 Objective test

In objective tests, the SNR is evaluated to quantify the overall quality of the enhanced speech signal. The SNR of the noisy signal is defined as

$$\text{SNR} = 10 \log_{10} \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} (x(n) - s(n))^2} \text{ dB} \quad (4.1)$$

The clean samples of a female speech (“She had your dark suit in greasy wash water all year”) and a male speech (“Should we chase those cowboys?”) signals are corrupted by additive white Gaussian noise for various SNRs ranging from  $-10$  to  $30$  dB. The noisy speech signal is then denoised using the proposed technique. The soft thresholding alone and the combination of hard and soft thresholding defined in the previous chapter are applied in both wavelet and DCT domain for enhancement of the speech signal. The average results of 25 independent runs for each SNR ( $-10, -5, \dots, 30$  dB) are shown in Fig. 4.2 for the female speech  $s_1$  and in Table 4.3 for the male speech  $s_2$ . For comparison,



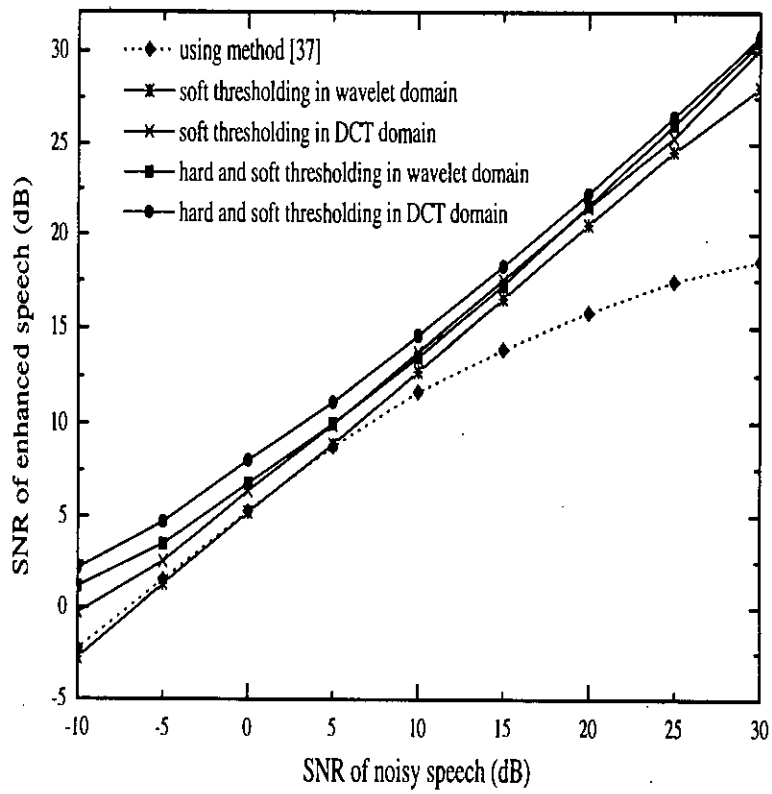


Fig. 4.2: Performance comparison in terms of input-output SNR for the speech, "She had your dark suit in greasy wash water all year", corrupted by additive white noise.

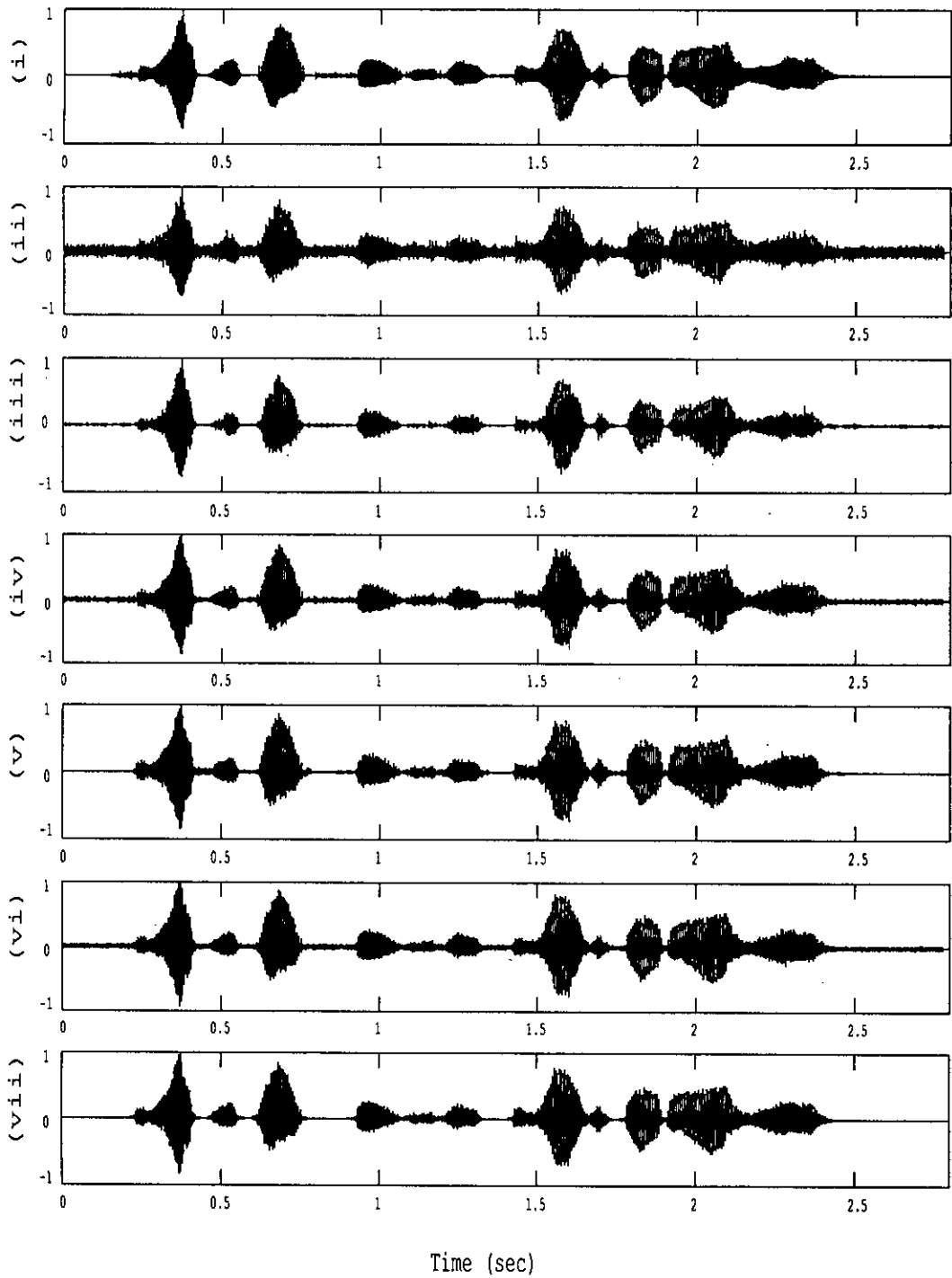
Table 4.4: Results on SNR improvement for two utterances (s3 and s4) corrupted by recorded real car noise

Speech	SNR (dB)	Wavelet packet using Ref. [37] (dB)	Proposed wavelet packet method		Proposed DCT method	
			soft thr. alone (dB)	hard and soft thr. (dB)	soft thr. alone (dB)	hard and soft thr. (dB)
s3	9.69	11.83	12.13	13.19	13.17	13.87
s4	17.93	19.17	19.91	20.93	20.76	21.83

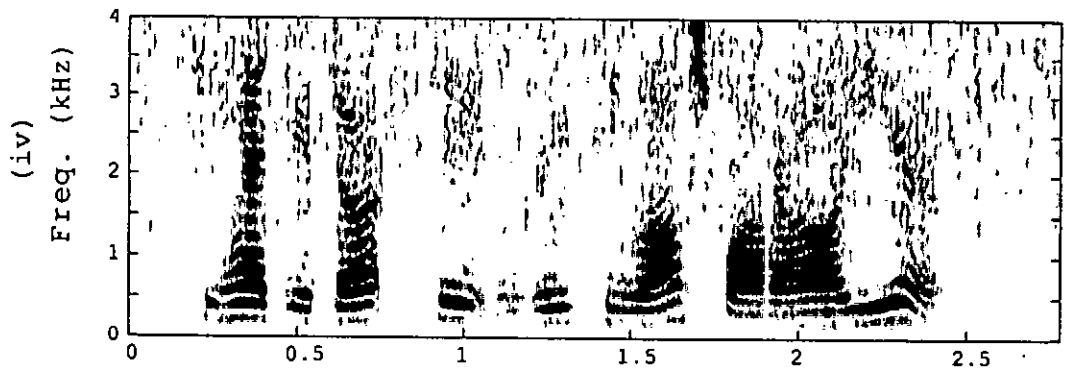
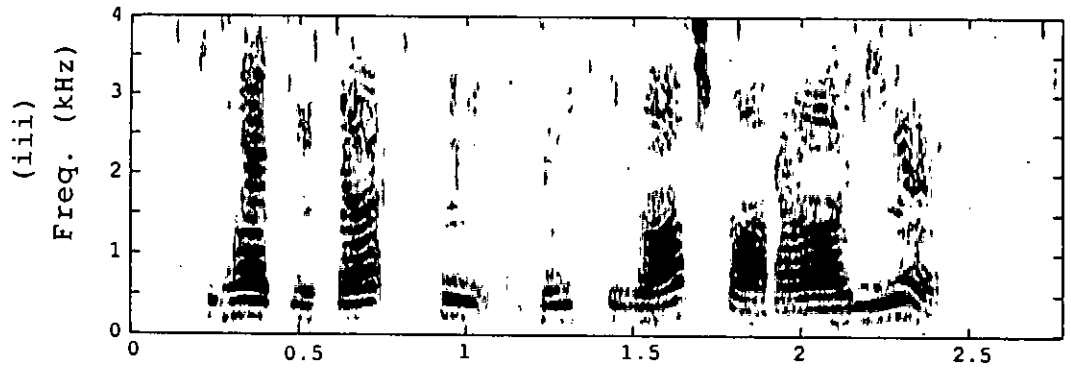
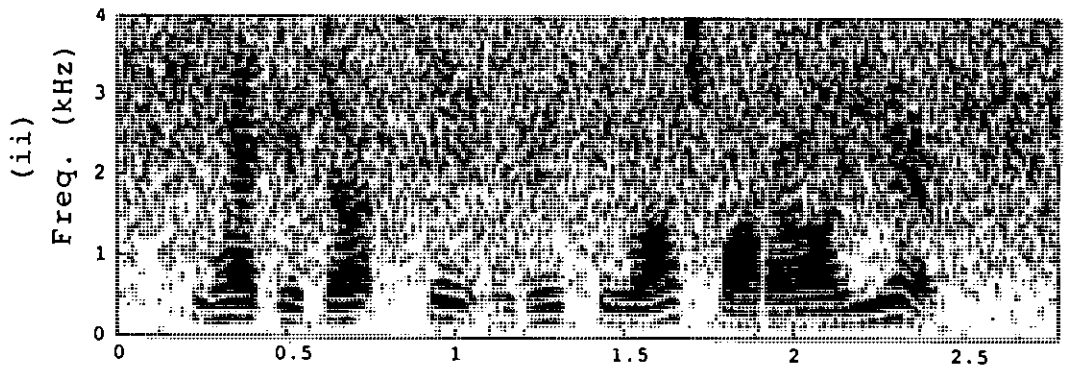
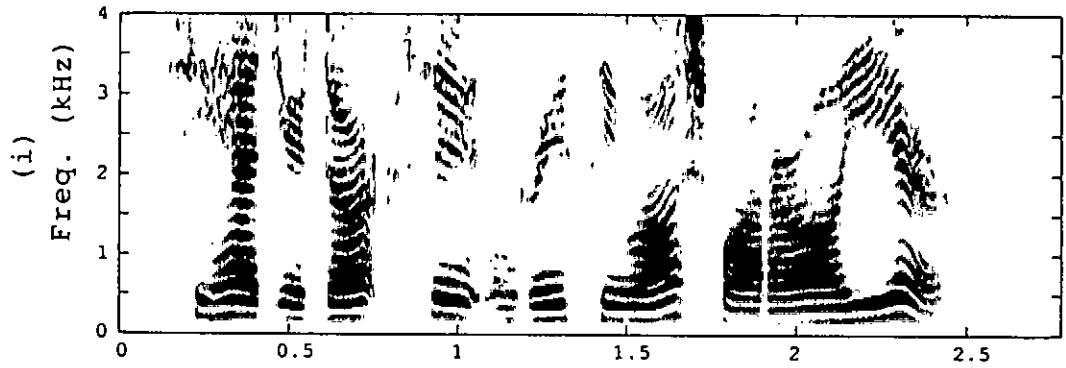
the results obtained using a recent method described in [37] are also included. It is evident that the proposed method in both wavelet and DCT domain show better enhancement performance than the previous one for almost all SNRs. The results for conventional soft thresholding in both wavelet and DCT domain also exhibit improved performance due to the compensation introduced in the biased noise level. Also notice that the proposed method prevents the undesired fall of SNR of the denoised speech even when the original signal has an SNR of 30 dB.

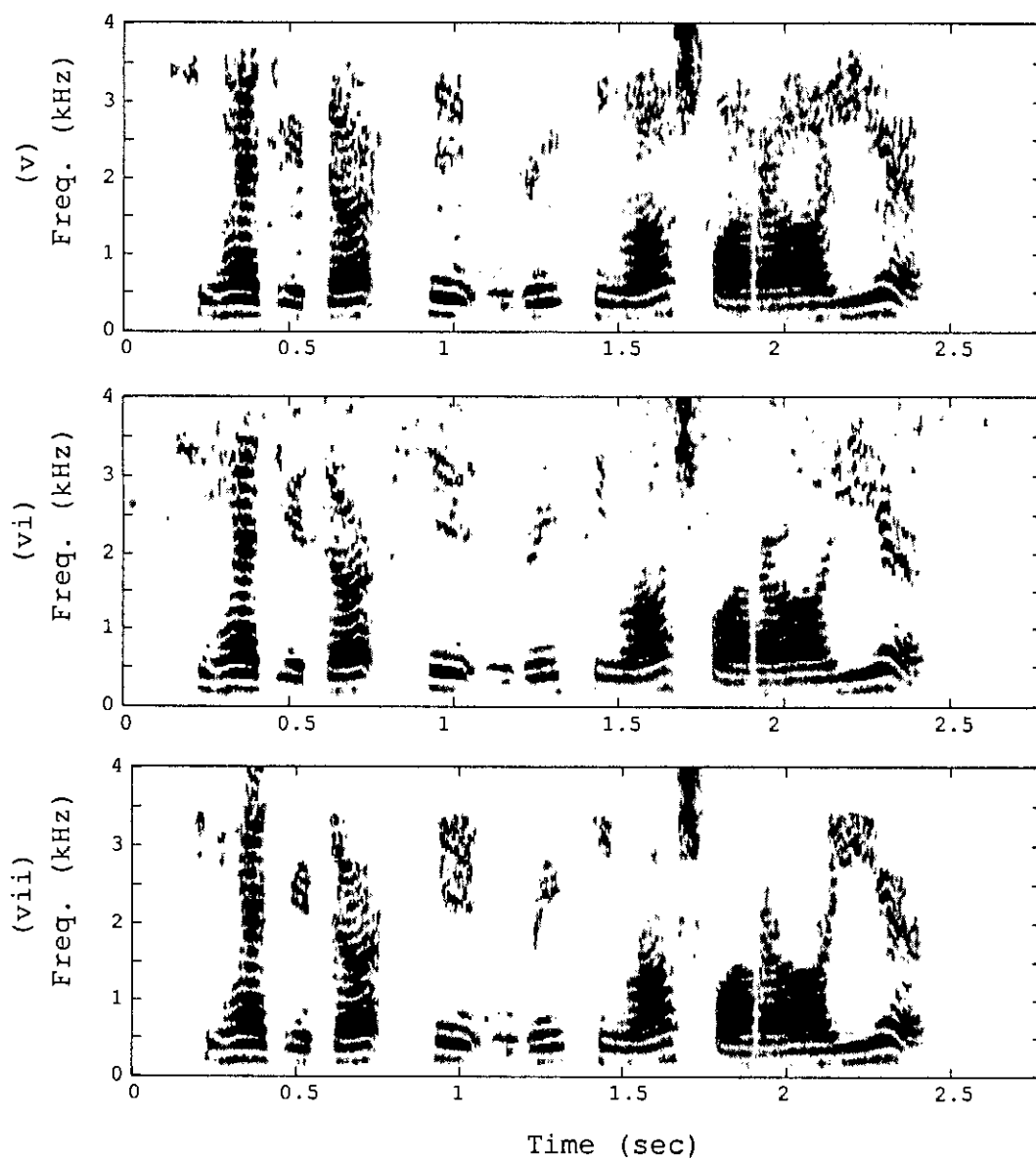
Fig. 4.3 (a) shows the white noise degraded speech  $x(n)$  at an SNR of 10 dB for the female speaker (s1) and the corresponding enhanced speech resulting from the wavelet packet method described in [37] and the denoising methods in wavelet and DCT domain using only soft and the combination of hard and soft thresholding proposed in this work are shown subsequently. The noise-free speech  $s(n)$  is also plotted along with the enhanced speech for comparison. It is apparent from Fig. 4.3 (a) that the proposed method eliminates noise in a better way than the one in [37]. It is also obvious from the spectrograms shown in Fig. 4.3 (b) that the enhanced speech by the proposed method includes less musical noise as compared to that of reported in [37]. Another speech (s2) by a male speaker is also corrupted by white noise and the simulation results in time domain are presented in Fig. 4.4 (a) and the corresponding spectrograms are in Fig. 4.4 (b).

Next, we generate noisy speech signal adding real noise recorded inside a slowly moving car with air cooler turned on with 17200 samples of a male and female clean speech, namely, "Would you please confirm government policy" termed as s3 and a numerical counting "1 2" termed as s4, respectively, from the TIMIT



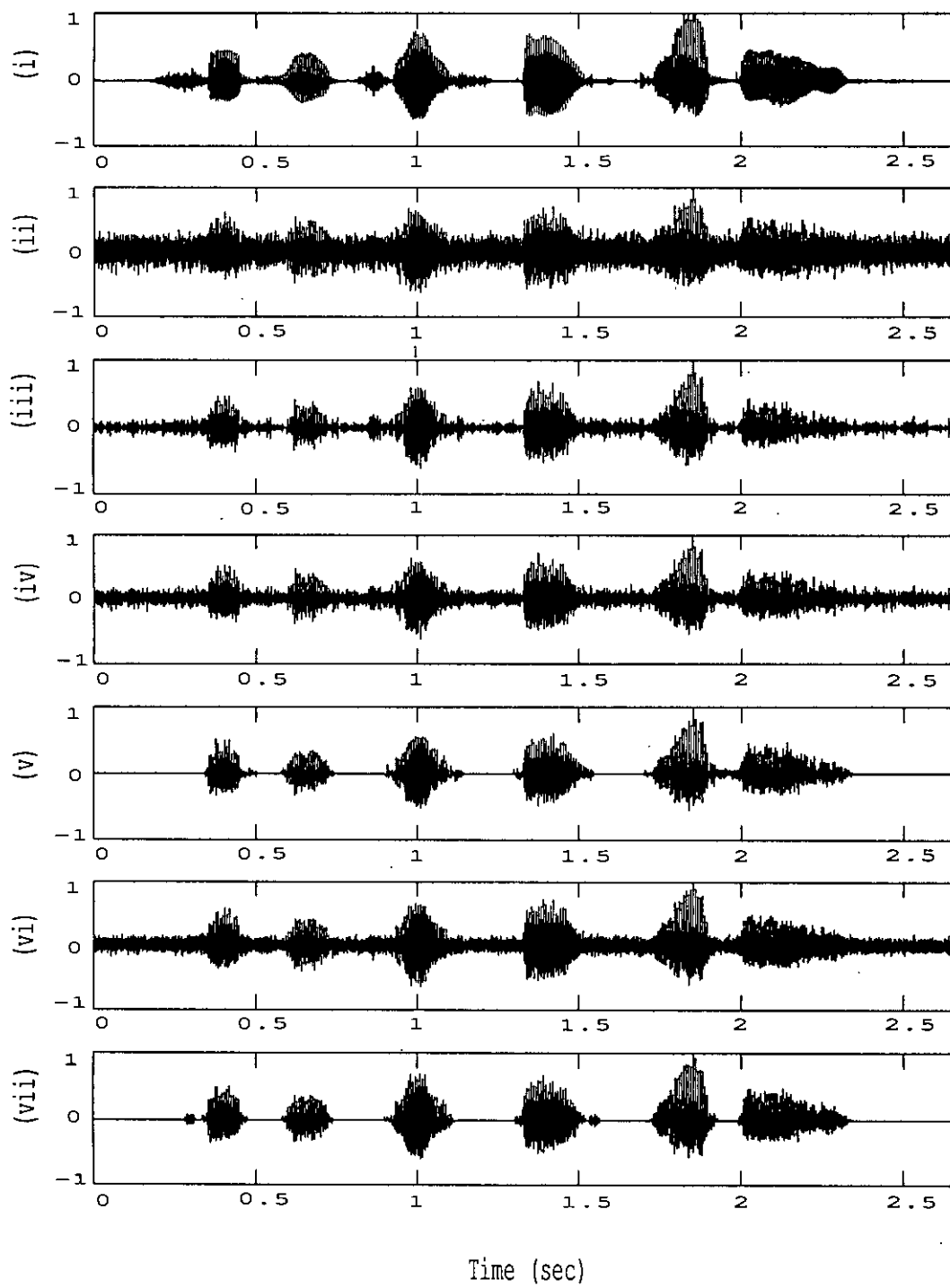
(a)



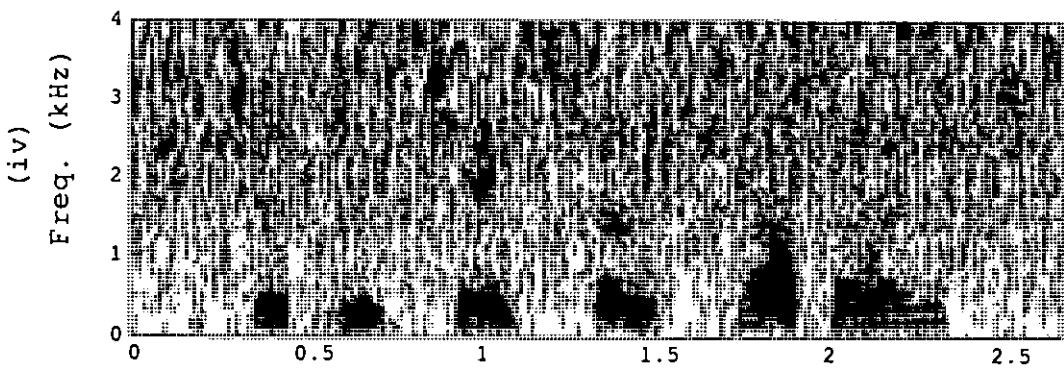
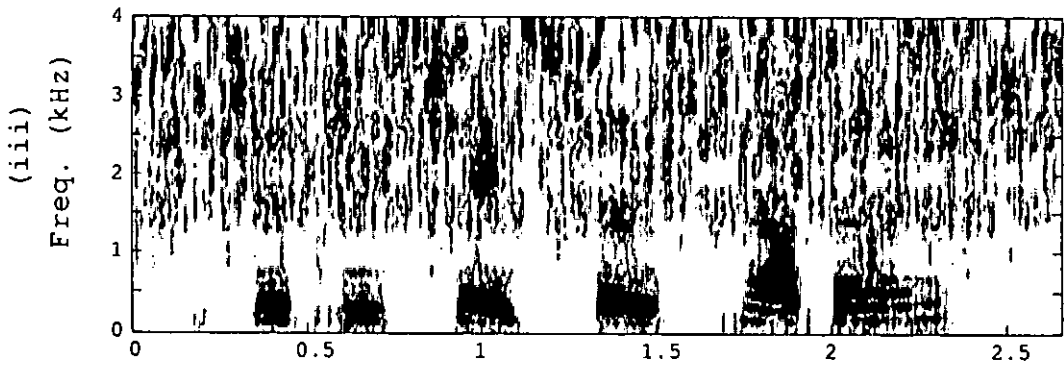
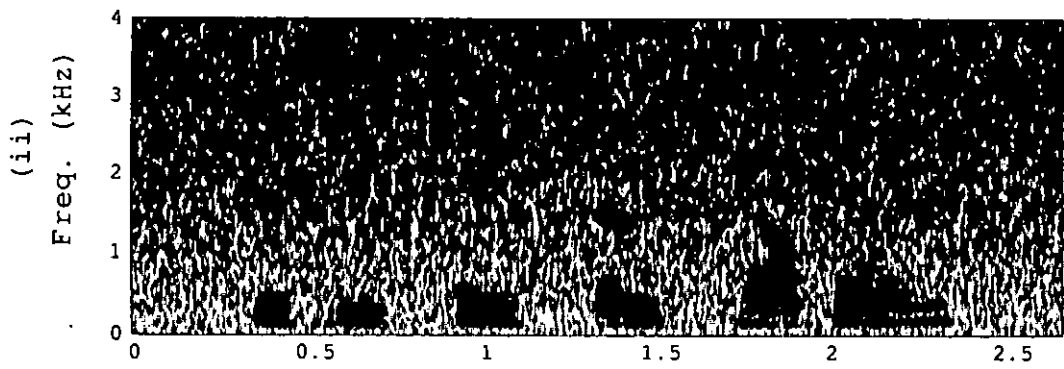
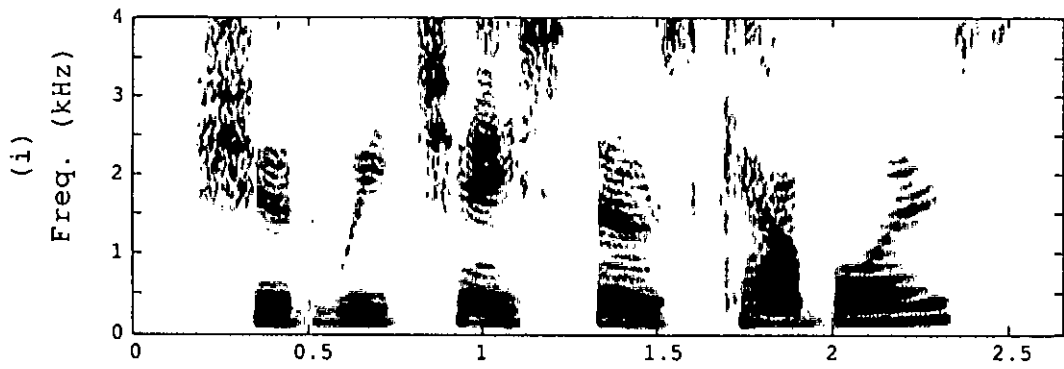


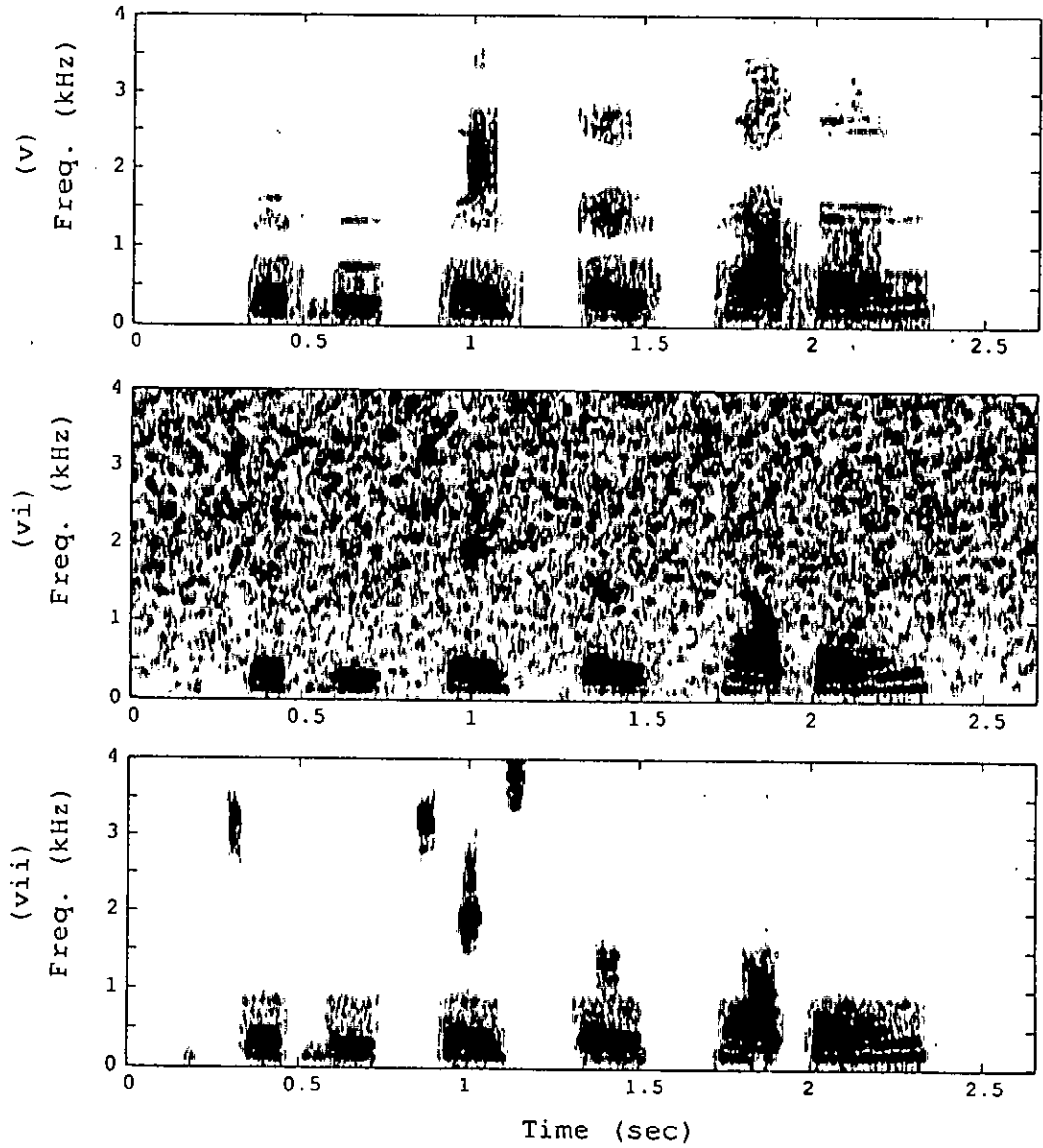
(b)

Fig. 4.3: Enhancement results for a female utterance “She had your dark suit in greasy wash water all year” corrupted by additive white noise: (a) Time-domain; (b) Spectrogram; (i) clean, (ii) noisy, (iii) denoised using Ref. [37], (iv) soft thresholding in wavelet domain, (v) hard and soft thresholding in wavelet domain, (vi) soft thresholding in DCT domain, (vii) hard and soft thresholding in DCT domain.



(a)

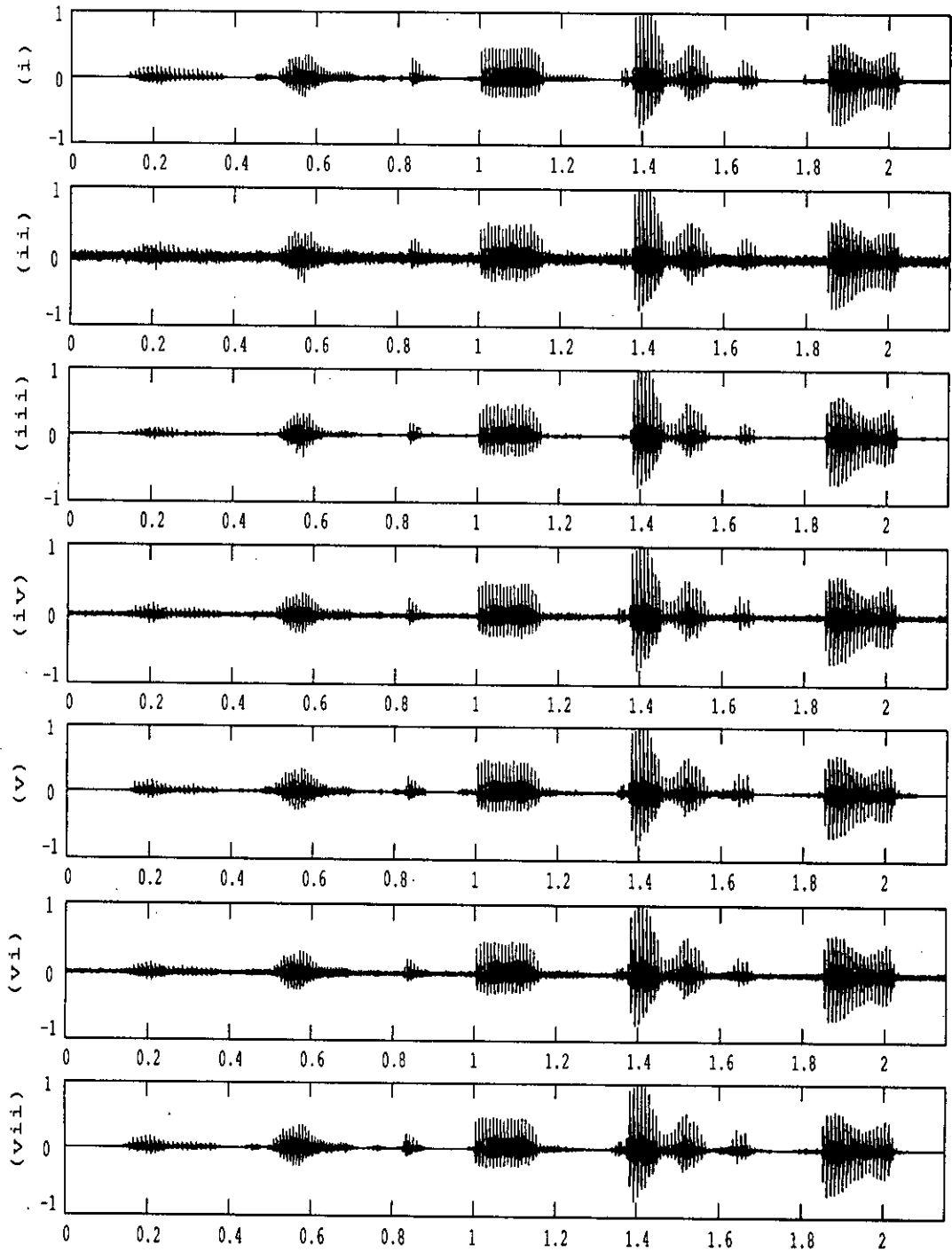




(b)

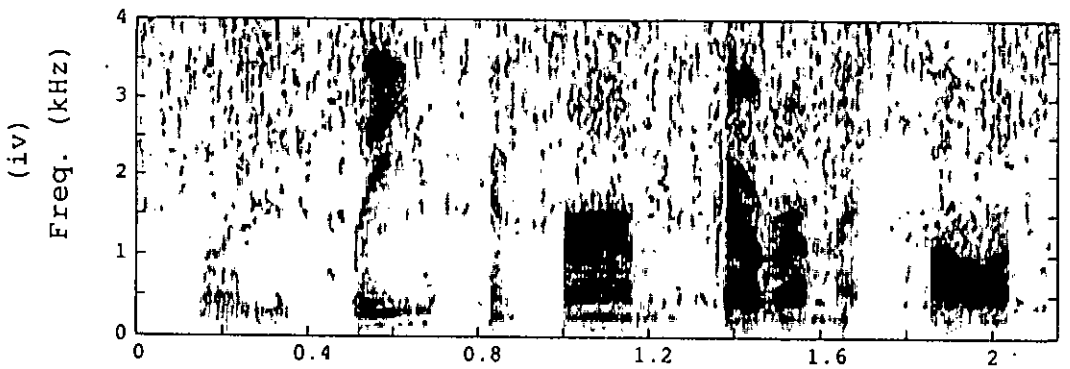
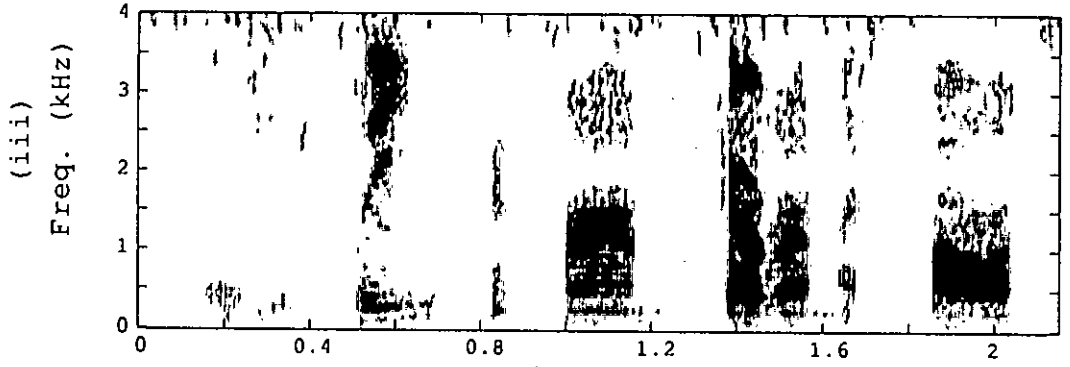
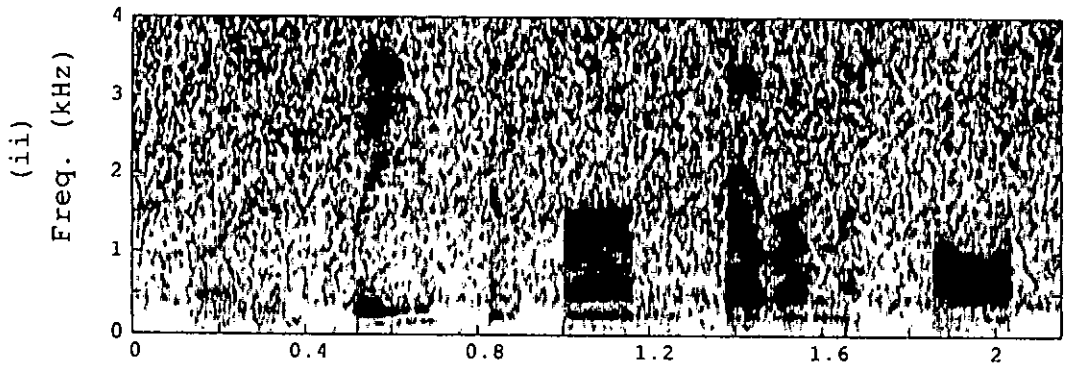
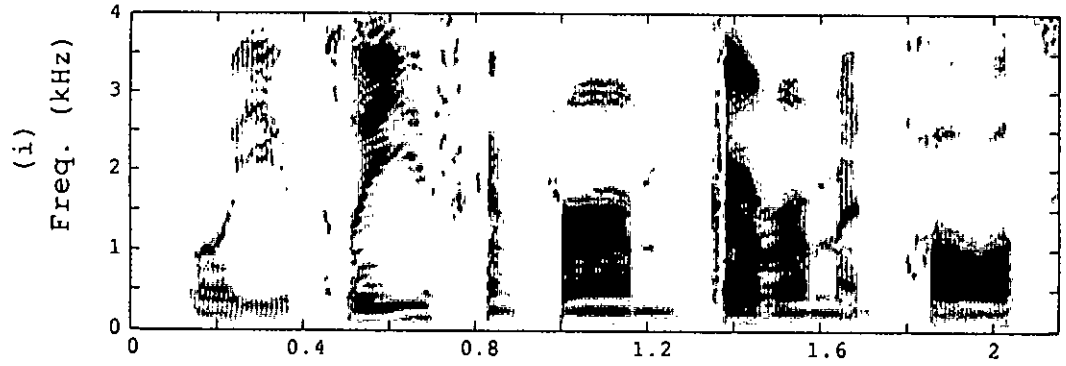
Fig. 4.4: Enhancement results for the utterance by a male speaker "Should we chase those cowboys?" corrupted by additive white noise: (a) Time-domain; (b) Spectrogram; (i) clean, (ii) noisy, (iii) denoised using Ref. [37], (iv) soft thresholding in wavelet domain, (v) hard and soft thresholding in wavelet domain, (vi) soft thresholding in DCT domain, (vii) hard and soft thresholding in DCT domain.

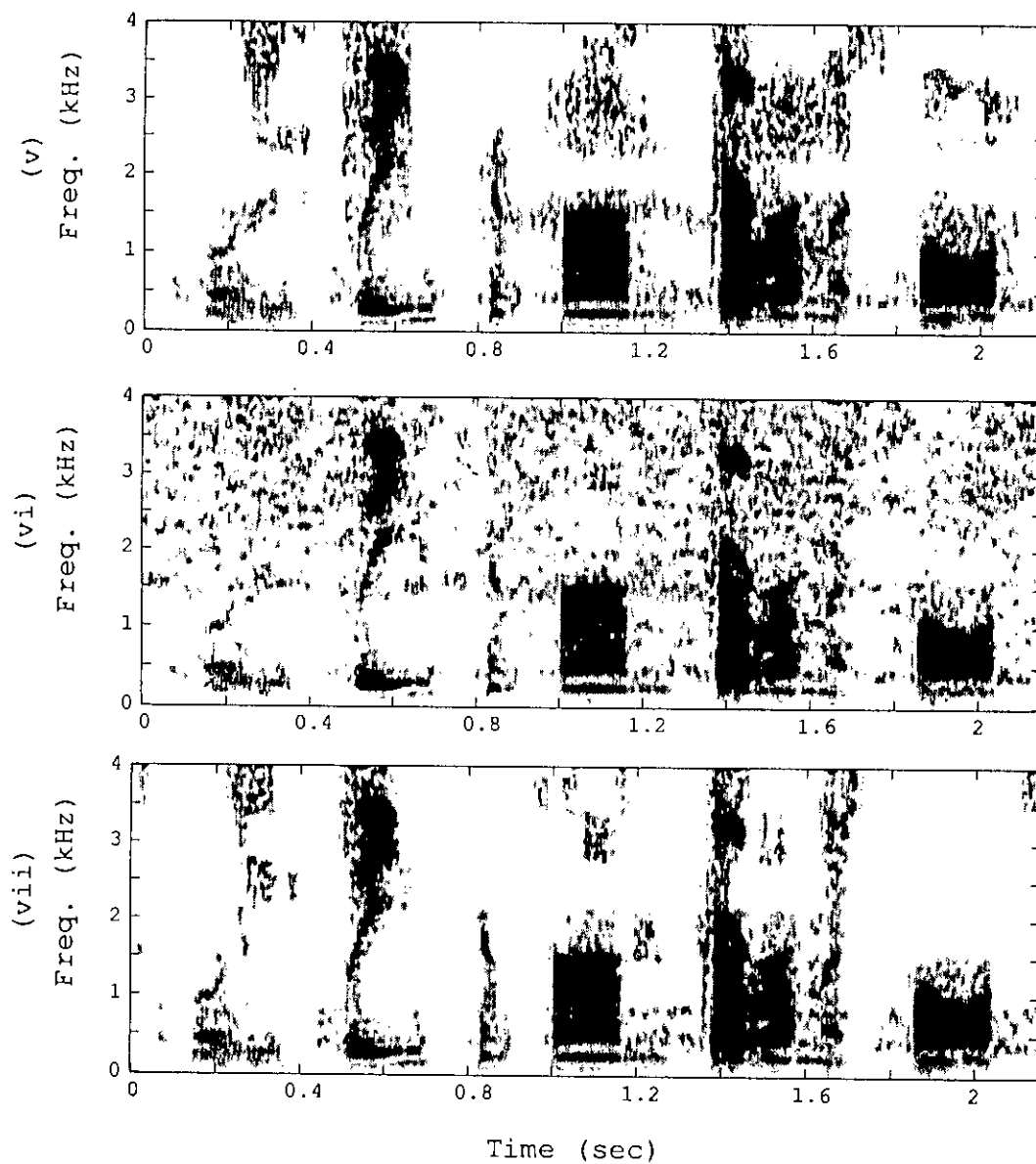




Time (sec)

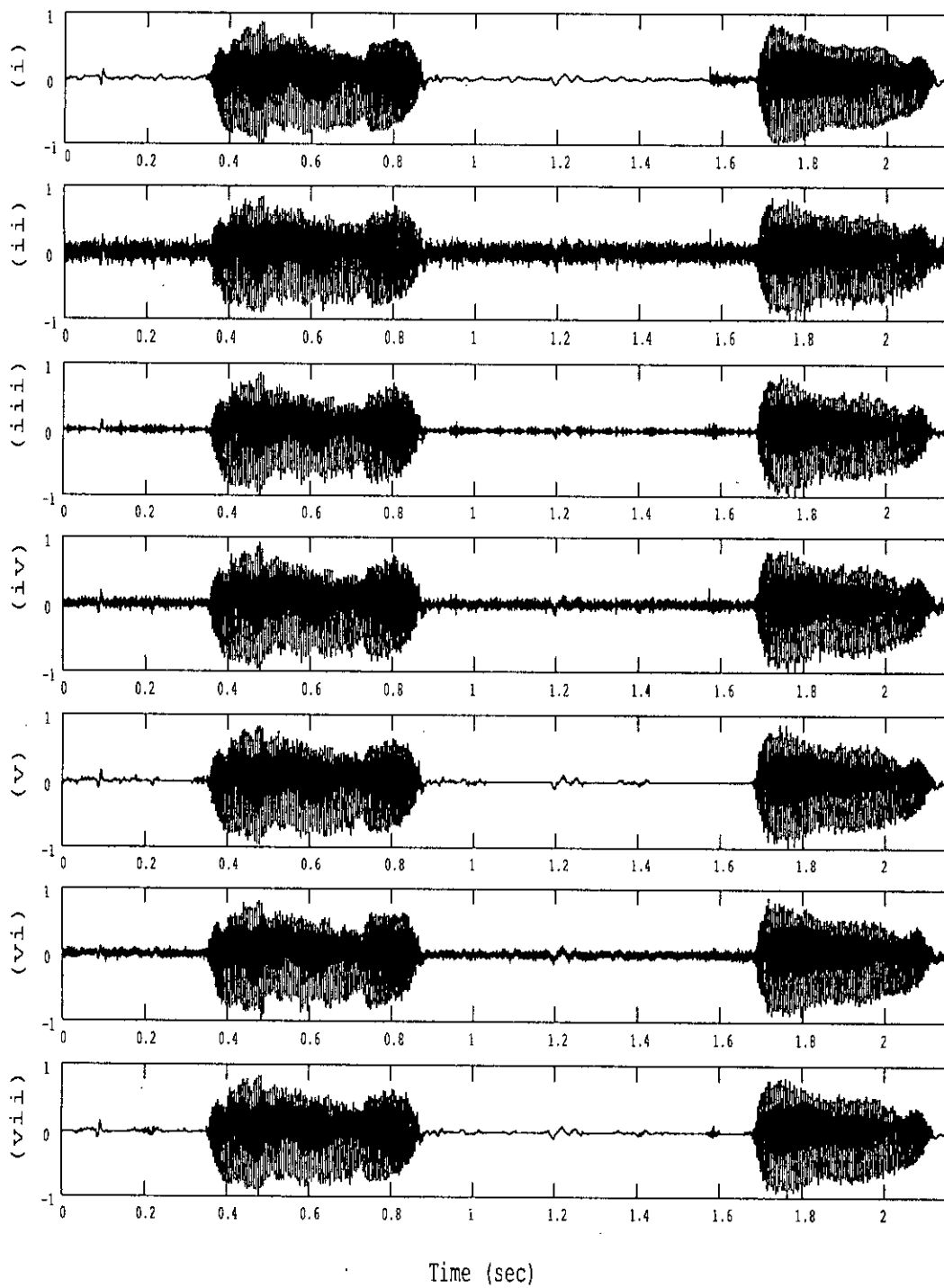
(a)



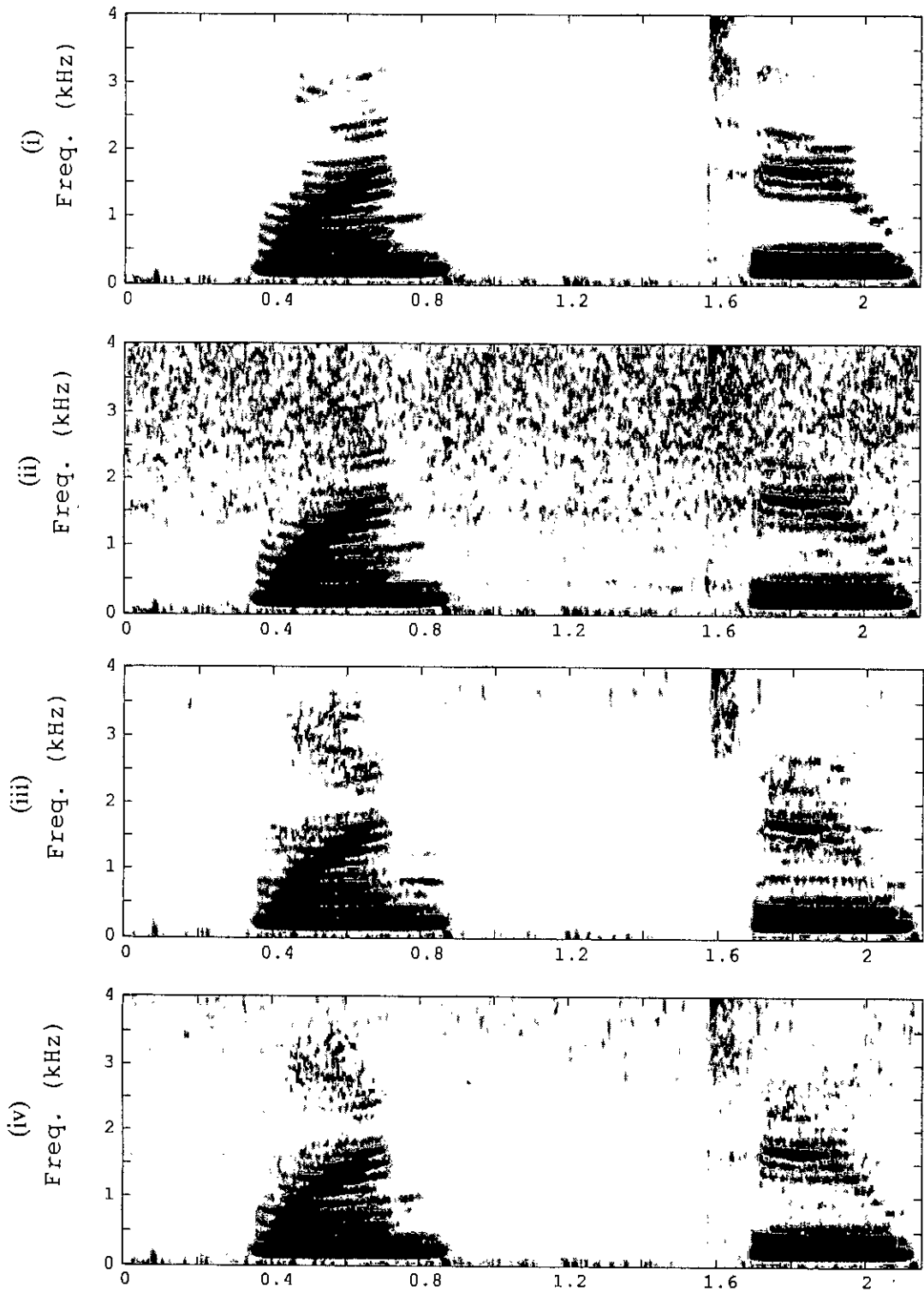


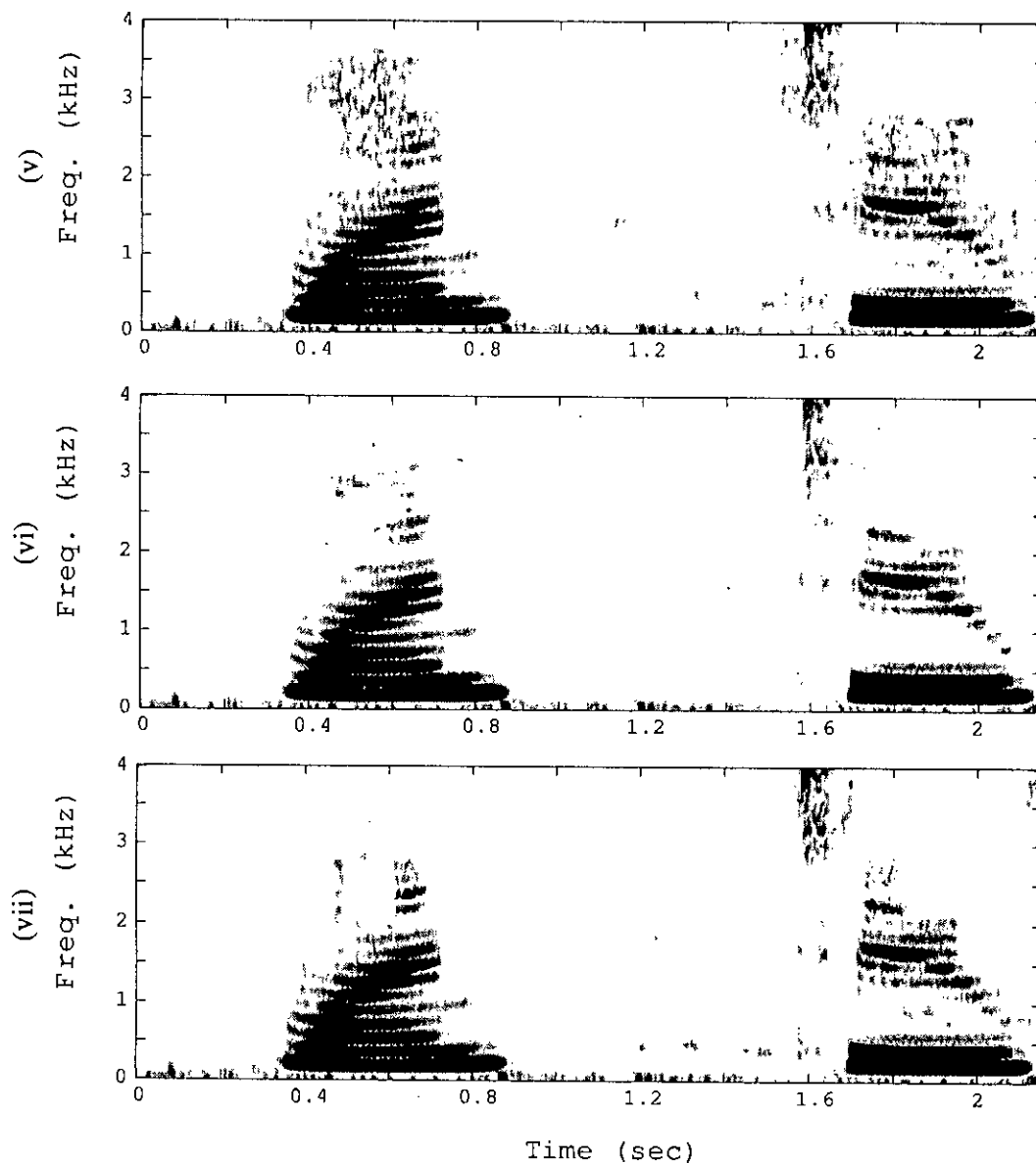
(b)

Fig. 4.5: Enhancement results for the utterance by a male speaker “Would you please confirm government policy” corrupted by real car noise: (a) Time-domain; (b) Spectrogram; (i) clean, (ii) noisy, (iii) denoised using Ref. [37], (iv) soft thresholding in wavelet domain, (v) hard and soft thresholding in wavelet domain, (vi) soft thresholding in DCT domain, (vii) hard and soft thresholding in DCT domain.



(a)





(b)

Fig. 4.6: Enhancement results for the numerical counting “One two” by a female speaker corrupted by the recorded real car noise: (a) Time-domain; (b) Spectrogram; (i) clean, (ii) noisy, (iii) denoised using Ref. [37], (iv) soft thresholding in wavelet domain, (v) hard and soft thresholding in wavelet domain, (vi) soft thresholding in DCT domain, (vii) hard and soft thresholding in DCT domain.

database to evaluate the performance of the proposed method. The SNR of the noisy speech was found to be 9.69 and 17.93 dB for s3 and s4, respectively. The corrected noise level is estimated using the similar procedure described in section 3.4.1. All other simulation conditions were kept same as that of white noise case. The SNR of the enhanced speech are shown in Table 4.4 for the proposed method along with the recent one [37]. It is also obvious that the proposed method in both wavelet and DCT domain show better enhancement performance for this recorded real noise. The enhancement results both in time and time-frequency domain are shown in Figs. 4.5 and 4.6 for s3 and s4, respectively. Also in this case, the proposed method removes noise comparatively introducing less distortion in the enhanced speech. This is also supported by the subjective test results presented in Tables 4.5 and 4.6.

### 4.3.2 Subjective test

It is known that SNR cannot faithfully indicate the quality of the enhanced speech. Thus we conduct subjected speech quality tests, employing a preference evaluation similar to one reported in [23]. The tests are performed by a group of 10 listeners with no previous familiarity with the test materials. The aforementioned two different speeches s1 and s3 are used in the test. Each subject participates in two listening sessions. In the first session, for each type of noise, listeners compare the outputs of the proposed system in the wavelet domain with the one reported in [37]. In the second session, the comparison is in between the outputs of the proposed system and the noisy input signal. In both sessions, listeners are asked to compare between a pair of speech signals played in random order and vote for one or none of them. Throughout the subjective tests, input SNRs are set to 5, 10, 15 and 20 dB. Table 4.5 summarizes the comparative results of both sessions.

In a similar fashion, another subjective test results between the proposed method in DCT domain and the one reported in [37] is given in Table 4.6. As shown, for both types of noise, the output of our proposed system has a higher preference percentage compared to the one reported in [37]. Also the listeners predominantly preferred the denoised signal using the proposed method as compared to the noisy signal.

Table 4.5: Results of subjective evaluation in terms of preference percentage in wavelet domain

Session	Noise type	Proposed method in wavelet domain	Wavelet method in [37]	Noisy signal	No preference
1st	white noise	40%	30%	-	30%
	real car noise	35%	20%	-	45%
2nd	white noise	90%	-	5%	5%
	real car noise	80%	-	10%	10%

Table 4.6: Results of subjective evaluation in terms of preference percentage in DCT domain

Session	Noise type	Proposed method in DCT domain	Wavelet method in [37]	Noisy signal	No preference
1st	white noise	65%	28%	-	7%
	real car noise	60%	28%	-	12%
2nd	white noise	80%	-	10%	10%
	real car noise	78%	-	12%	10%



## 4.4 Conclusion

In this chapter various results have been reported. The estimate of the bias-compensated noise levels are shown for two utterances by a male and a female speaker at different SNRs and the results are fairly close to the actual ones. As a result, SNR of the given noisy signal can be estimated accurately. The results for objective tests of the proposed method are shown along with the results of the recent method reported in [37]. Also the results for conventional amplitude subtraction based soft thresholding in both wavelet and DCT domain are produced using the corrected noise level as the threshold parameter. In addition to subjective evaluation, time and time-frequency plots of the proposed method in comparison to the recent one [37] are also reported at the end of this chapter. It can be conferred from the simulation results, i.e., from both subjective and objective tests, that DCT based method is preferable for speech enhancement over wavelet based method. Because, speech can be better represented by a sinusoidal model [51] than other basis functions usually used in the wavelet transform.

# Chapter 5

## Conclusion

### 5.1 Summary

A novel method for speech enhancement in wavelet and DCT domain has been proposed. The major focus of this research was to develop a highly accurate estimation of noise level considering the signal proportion remaining at the finest level. The upward bias due to signal proportion remaining at the finest level is reduced by exploiting the behaviour of the fourth-order statistics i.e., kurtosis of the transform noisy coefficients at the finest region. Unlike other conventional techniques, the signal-bias compensated noise level is then used as the threshold parameter for speech enhancement in both domain. Since coefficients at the trailing end contains both signal and noise component, neither hard nor soft thresholding alone is expected to result optimum enhancement. For this reason, a new thresholding technique is proposed that employs both hard and soft thresholding successively over the noisy transform coefficients. Hard thresholding is done by comparing the contribution of signal and noise strength to the restored speech. When the strength of noise is greater than signal, the coefficients are set to zero. Because, these coefficients contribute more noise than signal to the denoised speech. After hard thresholding, the power subtraction based soft thresholding is proposed to further reduce the noise in the enhanced speech. The results for conventional amplitude subtraction based soft thresholding in both wavelet and DCT domain are also produced to show the improvement over the recent one [37], which is only due to the correction introduced in the biased noise level. Using bias-compensated noise level as the threshold parameter, the proposed method show significant improvement in SNR than the very

recent results reported in [37]. The noise and signal power estimation schemes proposed here can be used for estimation of SNR with very good accuracy for further processing of the noisy speech signal.

## 5.2 Suggestions for future work

For noisy speech, energies of unvoiced segments are comparable to those of noise. Applying thresholding uniformly to all coefficients not only suppresses additional noise but also some speech components like unvoiced ones. Consequently, the perceptive quality of the filtered speech will be affected to some extent. The major goal of speech enhancement is to improve the perceptual aspects, i.e., intelligibility and quality of speech. The intelligibility of speech can be assured by applying the bias-compensated noise level as the threshold parameter. The proposed hard and soft thresholding has shown better performance for objective tests in terms of enhanced SNR. But, it may not always show better performance in subjective evaluations, due to the introduced distortions and artifacts known as the musical noise. Therefore, a modification in the thresholding technique may be investigated for further improvement of the quality of enhanced speech.

A linear assumption between the maximum value of cross-correlation (between template and test function),  $R_{max}$  and the correction factor,  $\beta$  has been adopted in this work. But practically it may have some deviation from linearity for different utterances. Therefore, a nonlinear approach incorporating speech dependent parameters may be proposed to get an exact correction factor for the biased noise level. Obviously, this will estimate the noise level more accurately and the SNR estimation scheme will be more appropriate.

Further research is needed to include other types of real noise with enhanced denoising performance.

# Bibliography

- [1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 113-120, Apr. 1979.
- [2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586-1604, Dec. 1979.
- [3] J. Makhoul *et al.*, *Removal of noise from noise-degraded speech signals*, Panel on removal of noise from a speech/noise signal, National Research council. Washington, DC : National academy press, 1989.
- [4] D. O'shaughnessy, "Enhancing speech degraded by additive noise or interfering speakers," *IEEE Commun. Mag.*, pp. 46-52, Feb. 1989.
- [5] S. F. Boll, "Speech enhancement in the 1980's: Noise suppression with pattern matching," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York : Marcel Dekker, 1992.
- [6] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. of IEEE*, vol. 80, no. 10, pp. 1526-1555, Oct. 1992.
- [7] I. Lecomte, M. Lever, J. Boudy and A. Tassy, "Car noise processing for speech input," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 512-515, May 1989.
- [8] N. D. Degan, C. Prati, "Performance of speech enhancement techniques for mobile radio terminal application," *Signal Processing III: Theories and applications*, New York: Elsevier Publishers B.V. (North Holland), pp. 381-385, 1986.

- [9] R. J. Niederjohn and J. H. Grotelueschen, "Speech intelligibility enhancement in a power generating noise environment," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, pp. 208-210, Aug. 1978.
- [10] I. Pollack, "Speech communications at high noise levels: The role of a noise operated automatic gain control system and hearing protection," *J. Acoust. Soc. Amer.*, vol. 29, pp. 1324-1327, Dec 1957.
- [11] I. B. Thomas and W. J. Ohley, "Intelligibility enhancement through spectral weighting," *Proc. IEEE Conf. Speech, Commun. and Processing*, pp. 360-363, 1972.
- [12] J. S. Lim and A. V. Oppenheim, "Reduction of quantization noise in PCM speech coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 107-110, Feb 1980.
- [13] Y. Ephraim and D. Malah, "Combined enhancement and adaptive transform coding of noisy speech," *Proc. Inst. Elec. Engg.*, vol. 133, pt. F, no. 1, pp. 81-86, Feb 1986.
- [14] O. Cappe, "Estimation of the musical noise phenomena with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 345-349, 1994.
- [15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Speech Audio Processing*, ASSP-32, pp. 1109-1121, 1984.
- [16] R. Hoeldrich and M. Lorber, "Broadband noise reduction based on spectral subtraction," *Proc. ICSPAT*, pp. 265-269, 1997.
- [17] P. Scalart and J. Vieira-Filho, "Speech enhancement based on apriori signal to noise estimation," *Proc. ICASSP*, pp. 629-632, 1996.
- [18] R. Mcaulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics Speech Signal Processing*, ASSP 28, pp. 137-145, 1980.

- [19] R. Mcaulay and M. Malpass, "Speech enhancement using a soft-decision maximum likelihood noise suppression filter," Tech. note 1979-31, M.I.T. Lincoln Lab., Lexington, MA, June 1979.
- [20] B. L. Sim, J. S. Chang, C. T. Tan and Y. C. Tong, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 328-337, 1998.
- [21] E. Nemer, R. Goubran, S. Mahmoud, "Speech Enhancement using fourth-order cumulants and optimum filters in the subband domain," *Speech Communication*, vol. 36, pp. 219-246, 2002.
- [22] C. Avendano, H. Hermansky, M. Vis and A. Bayya, "Adaptive speech enhancement using frequency specific SNR estimates," *Proceeding of III IEEE Workshop on Interactive voice Technology for Telecommunications Applications*, Basking Ridge, New Jersey, pp. 65-68, 1996.
- [23] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251-266, 1995.
- [24] J. Huang and Y. Zhao, "An energy-constrained signal subspace method for speech enhancement and recognition in white and colored noises," *Speech Communication*, vol. 26, pp. 165-181, 1998.
- [25] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov model," *IEEE Trans. Signal Processing*, vol. 40, pp. 725-735, 1992.
- [26] H. Sameti, H. Sheikhzadeh, L. Deng and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 445-455, 1998.
- [27] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, pp. 25-42, 1999.

- [28] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-26, pp. 197-210, 1978.
- [29] A. Erell and M. Weintraub, "Estimation of noise corrupted speech DFT-spectrum using the pitch period," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 1-8, 1994.
- [30] I. Y. Soon, S. N. Koh and C. K. Yeo, "Noisy Speech enhancement using discrete cosine transform," *Speech Communication*, vol. 24, pp. 249-257, 1998.
- [31] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 613-627, May 1995.
- [32] J. W. Seok and K. S. Bae, "Speech enhancement with reduction of noise components in the wavelet domain," in *Proc. of ICASSP*, pp. II-1323-1326, 1997.
- [33] D. Mahmoudi, "A microphone array for speech enhancement using multiresolution wavelet transform," in *Proc. Eurospeech'97*, Rhodes, Greece, pp. 339-342, 1997.
- [34] T. Gulzow, A. Engelsberg and U. Heute, "Comparison of a discrete wavelet transformation and nonuniform polyphase filter bank applied to spectral-subtraction speech enhancement," *Signal Process.*, vol. 64, pp. 5-19, 1998.
- [35] J. Sika and V. Davidek, "Multi-channel noise reduction using wavelet filter bank," in *Proc. Eurospeech'97*, Rhodes, Greece, 1997.
- [36] D. Mahmoudi and A. Drygajlo, "Combined Wiener and coherence filtering in wavelet domain for microphone array speech enhancement," in *ICASSP*, Seattle, WA, pp. 358-388, 1998.
- [37] M. Bahoura, J. Rouat, "Wavelet speech enhancement based on the teager energy operator," *IEEE Signal Processing Letters*, vol. 8, no. 1, pp. 10-12, January 2001.
- [38] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*, Englewood Cliffs, NJ: Prentice-Hall, 1978.

- [39] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 30, pp. 679-681, 1982.
- [40] M.R. Weiss, E. Aschkenasy and T. W. Parsons, "Processing of speech signals to attenuate interference," in *IEEE Symp. Speech Recognition*, (Pittsburgh, PA), pp. 292-293, 1974.
- [41] R. Le Bouquin, "Enhancement of noisy speech signals: Application to mobile radio communications," *Speech Communication*, vol. 18, pp. 3-19, 1996.
- [42] W. K. Pratt, *Digital Image Processing*, New York: Wiley, 1978.
- [43] L. R. Rabiner and R. W. Schafer, *Digital Processing of speech signals*, Englewood Cliffs, NJ: Prentice-Hall, 1978.
- [44] A. Gersho, R. M. Gray, *Vector Quantization and Signal Compression*, Boston: Kluwer Academic 1991.
- [45] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, pp. 257-286, Feb 1989.
- [46] J. C. Goswami, A. K. Chan, *Fundamentals of Wavelets: theory, algorithms and applications*, New York: John Wiley and Sons Inc., 1999.
- [47] J. F. Kaiser, "Some useful properties of Teager's energy operators," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, vol. 3, pp. 149-152, 1993.
- [48] D. L. Donoho, I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425-455, 1994.
- [49] Sir M. Kendall, A. Stuart, *The advanced theory of statistics*, Great Britain: Charles Griffin & Company Ltd., vol. 1, fourth edition, 1977.
- [50] John G. Proakis, Dimitris G. Manolakis, *Digital signal processing*, Prentice-Hall of India, 3rd edition, 2000.



- [51] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 7, pp. 731-740, Oct. 2001.

