

Speech Enhancement for White Noise Corrupted Speech Signals using Perturbation Technique

A Thesis Submitted to the
Department of Electrical and Electronic Engineering
of
Bangladesh University of Engineering and Technology

in Partial Fulfillment of the Requirement
for the Degree of
MASTER OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING

by
Muhammad Lutful Hai

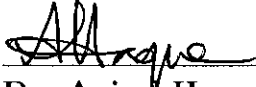


DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY


2005

The thesis titled “**Speech Enhancement for White Noise Corrupted Speech Signals using Perturbation Technique**” Submitted by Muhammad Lutful Hai, Roll No.: 100106264P, Session: October 2001 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of MASTER OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING on May 29, 2005.

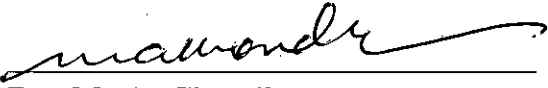
BOARD OF EXAMINERS

1. 

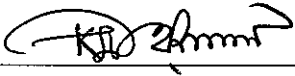
- Dr. Anisul Haque**
Professor
Department of Electrical and Electronic Engineering
BUET, Dhaka-1000, Bangladesh.
- Chairman**
(Supervisor)

2. 


- Dr. M. Rezwon Khan**
Professor and Vice-Chancellor (Designate)
United International University
Dhaka-1209, Bangladesh.
- Member**
(Co-supervisor)

3. 

- Dr. M. A. Choudhury**
Professor and Head
Department of Electrical and Electronic Engineering
BUET, Dhaka-1000, Bangladesh.
- Member**
(Ex-officio)

4. 

- Dr. Md. Kamrul Hasan**
Professor
Department of Electrical and Electronic Engineering
BUET, Dhaka-1000, Bangladesh.
- Member**

5. 

- Dr. Md. Abul Kashem Mia**
Professor
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 Lutful Hai

(Muhammad Lutful Hai)

Dedication

To My Family.

Contents

Declaration	iii
Dedication	iv
List of Abbreviations	x
Acknowledgement	xi
Abstract	xii
1 Introduction	1
1.1 Speech Enhancement: Background	1
1.2 Scope	3
1.3 Objective of the Work	5
1.4 Thesis Layout	6
2 Overview of Speech Enhancement Techniques	7
2.1 Introduction	7
2.2 Short Term Spectral Amplitude Techniques	8
2.2.1 Spectral Subtraction	9
2.2.2 Wiener Filtering	12
2.3 Adaptive Noise Canceling	13
2.4 Fundamental Frequency Tracking	13
2.5 Other Techniques	14
3 Speech Enhancement using Perturbation	15
3.1 Introduction	15

3.2	Amount of Perturbation in Frequency Domain	16
3.3	Estimation of Noise	19
3.4	Estimation of Noise Power ($P(v)$)	21
3.5	Conclusion	22
4	Simulation Results	23
4.1	Data Used for Simulation	23
4.2	Simulation Parameters	24
4.3	Identification of Noise Polarity	26
4.4	Performance Test	27
4.5	Objective Tests	31
4.5.1	Overall SNR	31
4.5.2	Average Segmental SNR	31
4.5.3	Itakura-Saito Distortion Measure	35
4.5.4	Spectrogram	39
4.6	Subjective Test	45
4.7	Conclusion	45
5	Conclusion	46
5.1	Summary	46
5.2	Limitations and Suggestion for Future Work	47
	Bibliography	48

List of Figures

2.1	Block diagram of Spectral Subtraction Algorithm	11
2.2	Flow diagram of ANC	14
3.1	Three sample frames of clean speech and its DCT	17
3.2	A frame of white gaussian noise and its DCT	18
3.3	Sample trend-lines of a frame	19
4.1	Characteristics of the factor Gamma (γ)	26
4.2	Polarity identification for different methods	28
4.3	Enhancement results for male utterance “Would you please confirm the government policy regarding waste removal”: (a) clean; (b) noisy (corrupted by 10dB WGN); enhanced using (c) MPE; (d) PARA; (e) MMSE-LSAE; (f) proposed Pert1 and (g) Proposed Pert2.	29
4.4	Enhancement results for female utterance “She had your dark suit in greasy wash water all year”: (a) clean; (b) noisy (corrupted by 10dB WGN); enhanced using (c) MPE; (d) PARA; (e) MMSE-LSAE; (f) proposed Pert1 and (g) Proposed Pert2.	30
4.5	Overall SNR for male utterance: Result is averaged using results of 20 utterances	33
4.6	Overall SNR for female utterance: result is averaged using results of 20 utterances	34
4.7	Average Segmental SNR improvement for male utterance: result is averaged using results of 20 utterances	37
4.8	Average Segmental SNR improvement for female utterance: result is averaged using results of 20 utterances	38
4.9	Itakura-Saito distortion measure for male utterance: result is averaged using results of 20 utterances	41

4.10 Itakura-Saito distortion measure for female utterance: result is averaged using results of 20 utterances	42
4.11 Spectrogram for a male utterance “Would you please confirm the government policy regarding waste removal”: (a) clean; (b) noisy (corrupted by 10dB WGN); enhanced using (c) MPE; (d) PARA; (e) MMSE-LSAE; (f) Proposed Pert1 and (g) Proposed Pert2 . . .	43
4.12 Spectrogram for a female utterance “She had your dark suit in greasy wash water all year”: (a) clean; (b) noisy (corrupted by 10dB WGN); enhanced using (c) MPE; (d) PARA; (e) MMSE-LSAE; (f) Proposed Pert1 and (g) Proposed Pert2	44

List of Tables

4.1	β values chosen for various SNR	25
4.2	Noise polarity identification	27
4.3	Overall SNR improvement for various noise levels, obtained using MPE, PARA, MMSE-LSAE and proposed two methods	32
4.4	Segmental SNR improvement for various noise levels, obtained using MPE, PARA, MMSE-LSAE and proposed two methods	36
4.5	Average Itakura-Saito Distortion Measure for various noise levels, obtained using MPE, PARA, MMSE-LSAE and proposed two methods	40

List of Abbreviations

ANC	Adaptive Noise Canceling
DFT	Discrete Fourier Transform
DSP	Digital Signal Processing
EM	Estimate Maximize
HMM	Hidden Markov Model
IS	Itakura-Saito
ITU	International Telecommunication Union
LMS	Least Mean Square
LP	Linear Prediction
MELP	Mixed Excited Linear Prediction
ML	Maximum Likelihood
MMSE	Minimum Mean Square Error
PDS	Power Density Spectrum
SNR	Signal to Noise Ratio
SOD	Second Order Difference
SS	Signal Subspace
STSA	Short Term Spectral Amplitude
WGN	White Gaussian Noise

Acknowledgement

I wish to convey my heartiest gratitude and profound respect to my co-supervisor Dr. M. Rezwan Khan and my supervisor Dr. Anisul Haque for their continuous guidance, suggestions and wholehearted supervision throughout the progress of this work, without which this thesis never be possible. Originally this research was supervised by Mr. Khan. The research was already in advanced stage when Prof. Khan left BUET joining United International University as Vice-Chancellor. I am grateful to him for acquainting me with the world of advanced research.

I am grateful to Mr. Abul Husain, Associate Professor and Head, Department of Electrical and Electronic Engineering, Ahsanullah University of Science and Technology, who provided with all the facilities of VLSI laboratory of the department and co-operation to complete the work. In this regard, I also like to express thanks and gratitude to Dr. M. A. Chaudhury, Professor and Head, Dept. of EEE, BUET.

I want to thank my friends Akeed Ahmed Pavel, Gour Das Podder and many others, who were directly or indirectly related to this work, for their support and encouragement.

Abstract

In this thesis a new technique is proposed to enhance speech degraded by noise. Speech enhancement is usually done mainly in two techniques, spectral subtraction and Wiener filtering. Both of the techniques use an attenuation filter without any concern of the polarity of noise with respect to clean speech. As a result conventional enhancement algorithm introduce additional distortion. In this work, a small perturbation is applied in DCT domain, effect of which is compared with the noisy observation in time domain to estimate the noise distribution in the same domain. A threshold level has been used for the application of perturbation. The noise so determined is used to enhance the signal. Computer generated White Gaussian Noise is used for simulation. The performance of the proposed technique is compared with three different methods proposed earlier. This promising new techniques shows better performance in terms of intelligibility for low signal-to-noise ratios when compared with the existing methods. Subjective tests also indicate better enhancement quality.

Chapter 1

Introduction

Speech enhancement is attracting a lot of researchers for long period of time. Still it is recognized as key topic in speech processing. In this chapter a brief introduction into the area of speech enhancement is presented, along with the objective and the way this thesis is poised.

1.1 Speech Enhancement: Background

When two persons communicate with each other in a quiet environment the information interchange between them is easy and accurate. But in a noisy environment listeners ability to understand the speech lessens. In many speech communication system background interference reduces the quality and intelligibility of the speech. Besides, originating from a noisy environment noise may be added while speech signal is being digitized or transmitted or received. Both digital and analog channels are possible, and communication can be either between people or with a machine. Speech enhancement assays to improve the performance of voice communication systems when the speech signal is corrupted by noise. The improvement is in the sense of minimizing the effects of noise while improving the quality and intelligibility of speech, particular to these systems. Examples of some important applications of speech enhancement are:

Telephone systems: In the telecommunication systems, the original speech is often corrupted by the ambient noise, e.g. in the cellular radio telephone systems, the original speech is corrupted by noise generated by the car engine [1, 2], traffic and wind as well as from competing voice (babble noise). Noise can also be introduced from the transmission channel [3]. The signals delivered by cellular systems may therefore be noisy with vitiated quality and intelligibility.

If the cellular system encodes the signal prior to its transmission, then further degradation in its performance results, since speech coders (vocoders) rely on some model for the clean signal and normally that model does not fit for the noisy signal. Recently, in the field of speech coding, considerable progress has been achieved in reducing the bit-rate while maintaining a high level of speech quality. Although vocoders, such as ITU G.729 and Mixed Excited Linear Prediction (MELP), give high quality for clean speech, it is significantly worse for coded noisy speech. Solution to circumvent this issue is to add a speech enhancement pre-processor that attenuates noise in the corrupted speech prior to encoding. 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 is unlikely to obey the statistical model for the clean signal used by the recognizer. Similar problems are encountered with pay phones located in noisy environments, such as airports, railway or bus station etc.; in hearing aids for people with hearing deficiency; in teleconferencing, where noise sources in one location may broadcast to all other locations.

Air to ground communication and vice-versa: In the air to ground communication systems, the cockpit noise corrupts the speech of the crew members. In such case, however, the messages of low quality and intelligibility delivered to the air traffic controllers may result disastrous effects. But communication from ground to air is rather simpler, since the noise is added to the speech at the channel and at the receiving end, respectively, rather than that at the source location. Hence, the clean signal can be immunized prior to being affected by the noise [4].

The aforementioned discussion corroborates that speech enhancement has three major goals: 1) improvement of perceptual aspects (e.g. *quality, intelligibility*) of a given sample function of degraded speech signal; 2) to increase the robustness of speech coders in presence of input noise; 3) to increase robustness of speech recognition systems in presence of input noise.

The *quality* of speech signal is a subjective measure, which reflects on the way the speech is perceived by a listener. It can be expressed in terms of how pleasant the signal sounds or how much easy a listener can understand the message lying in a sample speech. *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

the 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 for particular set of test condition (1. specific type of noise; 2. specific SNR; 3. noise estimate updates and 4. algorithm parameter settings), since there is no standard for quality assessment of different speech enhancement algorithm. Hansen et al proposed [5] a set of ingredients and a combination of objective measures and subjective testing rules to assess quality. However, researchers in the speech coding and recognition communities have standard criteria for algorithm performance comparison.

1.2 Scope

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 (babble noise), 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 or convolutional manner. Furthermore, the noise may be statistically dependent or independent (correlation between speech and noise) of the clean signal. The number of voice channels available for enhancements is an important factor in designing speech enhancement systems, In general, the larger the number of microphones, the easier the task of speech enhancement becomes. 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, speaker identifier), a digital communication system, etc.

Speech enhancement based on spectral decomposition and filtration [6, 7, 8, 9, 10, 11, 12, 13, 14, 15] remains a common and effective approach for enhancing speech degraded by acoustic additive noise when only the noisy speech is available. The relative lack of importance of phase for speech quality [16] has given rise to a family of speech enhancement algorithms based on spectral magnitude estimation. This general class is based on variations of optimum filters and en-

compasses such methods as spectral subtraction, Wiener filtering and various maximum likelihood (ML) estimation schemes. A common set of requirements in this class include: 1) An appropriate suppression rule based on an optimality criteria [9, 10] and which is usually a function of the SNR and other speech and noise statistics; 2) An estimation of the speech and noise power density spectrum (PDS), or their respective auto correlation; 3) A quantification of the probability of speech presence to further attenuate non-speech bands [11]; 4) A method for reducing residual noise by appropriately smoothing the estimated quantities [9] and/or exploiting the psychoacoustic properties of human hearing [17].

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., [10]) as well as formally derived ones. The ML estimation approaches in [9, 12] attempt to exploit the statistical properties of the discrete Fourier transform (DFT) of the 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 [9] whereby the variation in SNR between successive frames is reduced by averaging the locally computed SNR (SNR_{post}) with the SNR estimated in the previous frame after the filtering operation (SNR_{est}). The method results in a significant reduction of the ‘musical noise’ artifacts, as shown in [18].

Another speech enhancement approach is the signal subspace (SS) method [19, 20]. 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.

Considerable interest has been shown in recent years regarding wavelet as a new transform technique for both speech and image processing applications. Donoho introduced [21] wavelet thresholding (shrinking) as a powerful tool in denoising signals degraded by additive white noise. Although the application of wavelet shrinking for speech enhancement has been reported in several works (for example [22, 23]), there are many problems yet to be resolved for a successful application of the method to speech signals degraded by various noise types. Hidden Markov Model (HMM) based speech enhancement approaches [24, 25, 26]

and also combination of both Wavelet and HMM based enhancement techniques [27] have drawn much attention in recent years.

Methods for speech enhancement have also been developed based on extraction of parameters from noisy speech, and then synthesizing speech from these parameters [28]. All-pole modeling of degraded speech is one such method [29]. 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 [30]. 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 articular. These methods for enhancement of speech depend critically on the estimation of pitch from the noisy speech signal.

Many speech enhancement algorithms use DFT as the transform domain for removing noise embedded in the noisy speech signal [1]- [30]. Wavelet transform is also used as the transform domain as mentioned earlier. Recently, discrete cosine transform (DCT) has been widely used as the analysis tool in the field of speech enhancement [31, 32, 33, 34, 35], which is well accepted now in speech processing because its coefficients are real and the noise components can add or subtract with the actual signal coefficients and for its higher energy compaction.

1.3 Objective of the Work

The objective of this research is to introduce a new technique in speech enhancement which capitalizes on perturbing a signal in a transform domain and observing the corresponding change in time domain to reach a conclusion about the actual polarity as well as the distribution of noise in a noisy speech signal. Conventional speech enhancement techniques estimate a noise bias and subtract that noise bias from the corresponding noisy signal spectrum (known as spectral subtraction, e.g. [6, 8]) or multiplies the noisy signal by a gain factor to have an estimate of clean speech (known as Wiener filter, e.g. [7, 9]). In most of the cases relative polarity of noise with respect to clean signal has always been ignored because of the assumed lack of importance of phase in speech enhancement [16].

DCT is chosen as the transform domain for this work. A perturbation in the form of small change of estimated noise level is applied in DCT domain.

The corresponding changes in time domain of the perturbed signal is used to identify the noise distribution in time domain. A filtration is applied based on the obtained noise distribution. A good SNR as well as good quality is observed for the processed speech.

1.4 Thesis Layout

This thesis consists of five chapters. Chapter 1 gives a brief description of necessity of speech enhancement techniques, names of existing methods and the main objectives of this research work.

In chapter 2, a brief review of existing various speech enhancement techniques such as spectral subtraction rules, Wiener filtering and Wavelet based enhancement techniques are presented.

In chapter 3, the drawback of the conventional speech enhancement techniques is discussed. To improve their performances, a new technique is proposed using perturbation technique.

The simulation results for proposed algorithm is presented in chapter 4. The proposed algorithm is compared with the results proposed by Hasan *et al.* [35] and by Ephraim *et al.* [36]. Both subjective and objective evaluations are also reported along with necessary measurements in this chapter.

In chapter 5, the thesis is concluded by presenting an overall discussion on this research and pointing out some drawbacks of the proposed algorithm along with some suggestion for future work.

Chapter 2

Overview of Speech

Enhancement Techniques

2.1 Introduction

There are a number of ways in which speech enhancement systems can be classified. A broad grouping is concerned with the manner in which the speech is modelled. Some methods are based on stochastic process models of speech which rely on a mathematical criterion while others are based on perceptual aspects of speech that attempts to improve aspects important to human perception.

Enhancement algorithms can also be classified depending on whether a single-channel or dual-channel (multichannel) approach is used. In dual-channel system one channel (consist of a microphone or a set of microphones, known as primary channel) is used to receive noisy speech (clean speech plus noise) and another channel (a microphone or a set known as secondary channel) is used to contain a sample of noise correlated to the noise in primary channel. For single-channel systems only one microphone is available, so noise sample is extracted during the period of silence.

Beyond the classification based on specific aspects of speech, there are four broad classes of enhancement that differ substantially in the general approaches. Success of the classes depends on their own set of assumptions, also have specific advantages and limitations.

First type concentrates on short-term spectral amplitude (STSA) where an estimation of noise bias is subtracted from noisy signal in power spectral, fourier domain or in autocorrelation domain. The estimated noise bias is

calculated from non-speech pause intervals for a single-channel system or from a reference microphone for multi-channel system.

Second type is based on speech modeling using iterative methods, also widely known as Wiener Filtering. These systems focus on estimating model parameters that characterize the speech signal, followed by resynthesis of the noise-free signal based noncausal. This techniques requires *a priori* knowledge of noise and speech.

Third type is based on adaptive noise canceling (ANC) that is formulated using a dual-channel time or frequency domain environment based on the least mean square (LMS) algorithm.

Fourth type is based on the periodicity of voiced speech. These methods employ fundamental frequency tracking using either single channel ANC or adaptive comb filtering of the harmonic magnitude spectrum.

Below the above mentioned techniques will be discussed very briefly.

2.2 Short Term Spectral Amplitude Techniques

This type of speech enhancement techniques based on processing in short-term spectral domain. It is one of the earliest and perhaps the easiest to implement. In this family of methods an estimate of the spectral amplitude associated with the original signal is obtained without considering the phase of original signal. This is because spectral amplitude rather than the phase is important for speech quality and intelligibility. A variety of speech enhancement techniques capitalize on this aspect of speech perception by focusing on enhancing only the spectral amplitude. The techniques to be discussed can be broadly classified into two groups. **First**, the enhancement procedure is performed over frames by obtaining the short-term magnitude and phase of the noisy spectrum, subtracting an estimated noise magnitude spectrum from the noisy speech magnitude spectrum, and inverse transforming this spectral amplitude using the phase of the original degraded speech [6, 8, 37]. **Second**, the degraded speech is first used to obtain a filter (Wiener filter) which is then applied to the degraded speech. Since these procedures lead to zero-phase filters, here again only the spectral amplitude is

enhanced, with the phase of the filter being the same to that of the degraded speech [29].

2.2.1 Spectral Subtraction

The spectral subtraction method was first proposed by Boll [6]. An estimate of STSA is obtained by subtracting the estimate of the noise spectrum from the noisy speech spectrum. Information about the noise spectrum is obtained during non-speech activity. After finding the spectral estimator, spectral error is computed and some additional modifications are made for reducing it.

In spectral subtraction analysis some important assumptions are made:

1. The background noise is added acoustically or computationally to the speech.
2. The background noise environment remains locally stationary to the degree that spectral magnitude expected value prior to speech activity equals its expected value during speech activity.
3. Significant noise reduction is possible by removing the effect of noise from magnitude spectrum.

Let $s(n)$, $d(n)$ and $x(n)$ be representing clean speech, noise and noisy speech, respectively. The noise $d(n)$ is assumed to be uncorrelated, i.e. the autocorrelation function of $d(n)$ be:

$$r_d(\eta) = D_0\delta(\eta) \quad (2.1)$$

Where D_0 is some constant, $\delta(\eta)$ is the impulse sample at η and η is the autocorrelation lag. Realizations $s(n)$, $d(n)$ and $x(n)$ are related by:

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

As $d(n)$ is an uncorrelated process, it follows:

$$\Gamma_x(\omega) = \Gamma_s(\omega) + \Gamma_d(\omega) \quad (2.3)$$

Where $\Gamma(\cdot)$ denotes the PDS. PDS of any signal, say $y(n)$ is defined as:

$$\Gamma_y(\omega) = \frac{1}{N} \sum_{\eta=-\infty}^{\infty} r_y(\eta)e^{-j\omega\eta} \quad (2.4)$$

Where $r_y(\eta)$ is the auto-correlation function.

If an estimate of $\Gamma_d(\omega)$, say $\hat{\Gamma}_d(\omega)$ is obtained then it is possible to estimate the PDS of uncorrupted speech as:

$$\hat{\Gamma}_s(\omega) = \Gamma_x(\omega) - \hat{\Gamma}_d(\omega)$$

or, equivalently,

$$|\hat{S}(\omega)|^2 = |X(\omega)|^2 - |\hat{D}(\omega)|^2$$

Here $|\hat{S}(\omega)|$, $|X(\omega)|$ and $|\hat{D}(\omega)|$ are the magnitude spectrum of estimated clean signal, noisy signal and estimated noise respectively.

At a given frequency, the estimated amplitude of the signal is the combined amplitudes of the pure signal and noise, minus the estimated noise amplitude. The phase of the enhanced signal is assumed to be the same as that of the noisy signal. Then estimated enhanced signal $\hat{S}(\omega)$ be

$$\begin{aligned} \hat{S}(\omega) &= |\hat{S}(\omega)|e^{j\phi_x(\omega)} \\ &= \left[|X(\omega)|^2 - |\hat{D}(\omega)|^2 \right]^{\frac{1}{2}} e^{j\phi_x(\omega)} \end{aligned} \quad (2.5)$$

Where $\phi_x(\omega)$ is the phase function. This modification of the frequency domain information can be placed within the overlap and add framework to regenerate a time signal with better signal-to-noise ratio. In other words, the estimate of enhanced signal can be written as the sum of the signal and the non-stationary component of noise.

Therefore the spectral error is:

$$\epsilon(\omega) = \hat{S}(\omega) - S(\omega) \quad (2.6)$$

This is the formulation behind spectral subtraction method. The block diagram of spectral subtraction algorithm is given in Fig. 2.1.

In addition to the formulation given above Boll proposed a number of simple modifications to reduce the effects of spectral error of the estimated signal [6]. These are:

1. *Magnitude averaging*: Since the spectral error equals the difference between the noise spectrum and its expected value, averaging local spectral magnitudes can reduce the error. Hence the sample mean of the noise magnitude spectrum will converge to its mean as longer averages are taken. The magnitude-averaged spectrum is found using the sample mean

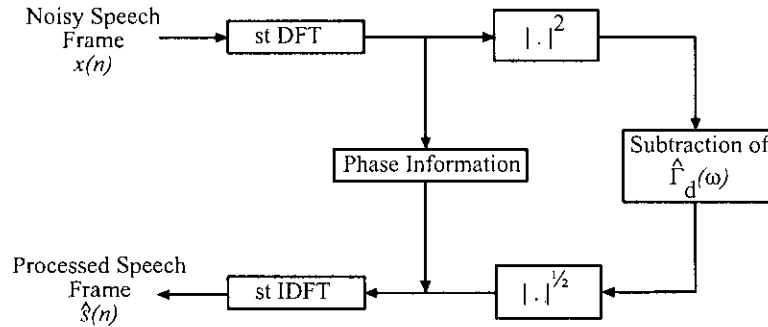


Figure 2.1: Block diagram of Spectral Subtraction Algorithm

$$\overline{|X(\omega)|} = \frac{1}{2I+1} \sum_{l=i-I}^{i+I} |X_l(\omega)|$$

Where X_l is the time-windowed transform of the frame centered in $2I + 1$ frames. Then the resultant estimator for Eq. (2.5) will be

$$\hat{S}(\omega) = \left[\overline{|X(\omega)|} - |\hat{D}(\omega)| \right] e^{j\phi_x(\omega)} \quad (2.7)$$

The obvious problem with this modification is that the speech is non-stationary, and therefore only limited time averaging is allowed. The major disadvantage of averaging is the risk of some temporal smearing of short transitory sounds.

2. *Half wave rectification:* For each frequency value where noisy speech magnitude spectrum is less than average noise spectrum, the output is set to zero. Half-wave rectification is generally employed in spectral subtraction methods to avoid the impossible case of negative energy values. The advantage of this procedure is that noise floor is reduced by the average magnitude of the noise spectrum and low variance coherent noise tones are eliminated. However we can sometimes observe the incorrect removal of speech information which can be counted as a disadvantage of half-wave rectifying.
3. *Residual noise reduction:* After half-wave rectification speech & noise lying above the expected value of noise magnitude spectrum remains. During non-speech activity this noise residual will exhibit itself with magnitude

between zero and a maximum value measured during non-speech activity. Since the noise residual has audible effects in the time domain (it will sound like the sum of tone generators with random frequencies), it should be reduced by replacing its current value with its minimum value chosen from adjacent analysis frames. There are three reasons for using this procedure:

- If the amplitude of spectral subtraction estimator lies below maximum noise residual and fluctuates then the spectrum at that frequency is most probably due to noise.
- If the amplitude of spectral subtraction estimator lies below maximum noise residual but has a constant value then the spectrum at that frequency is most probably low energy speech.
- If the amplitude of spectral subtraction estimator lies above maximum noise, there is speech present at that frequency.

Numerous researchers proposed various modification over the above mentioned methodology. Some of such modifications can be found in [8] by Berouti *et al* (1979), [37] by McAulay and Malpass (1979).

2.2.2 Wiener Filtering

In Wiener filtering method a frequency weighting for an optimum filter is first estimated from the noisy speech signal $x(n)$. This filter (say $h(n)$) is *linear* and *noncausal* and is then applied to the noisy speech to obtain an estimate of $s(n)$, say $\hat{s}(n)$. The optimality of the filter is in MMSE sense, that is the impulse response $h(n)$ be such that

$$\xi = E\left\{[s(n) - \hat{s}(n)]^2\right\} \quad (2.8)$$

is minimized. Using the orthogonality principle the Wiener filter in the frequency domain is found to be

$$H(\omega) = \left[\frac{\Gamma_s(\omega)}{\Gamma_s(\omega) + \Gamma_d(\omega)} \right]$$

where $\Gamma_s(\omega)$ and $\Gamma_d(\omega)$ are the PDS of clean speech $s(n)$ and noise $d(n)$ respectively.

In practice this filter cannot be obtained, since speech signal $s(n)$ is only short-term stationary and also spectrum of $s(n)$ (PDS) is not known. One way

to overcome this to use an approximate filter with frequency response

$$H(\omega) = \left[\frac{\hat{\Gamma}_s(\omega)}{\hat{\Gamma}_s(\omega) + \hat{\Gamma}_d(\omega)} \right] \quad (2.9)$$

where hats indicates that those quantities are estimated. In Eq.(2.9) $\hat{\Gamma}_d(\omega)$ can either be found from an assumed known statistics of noise or may be calculated from the pause intervals as noise can be assumed to be stationary. But $\hat{\Gamma}_s(\omega)$ is not so easy to obtain. One approach to estimate the speech spectrum is to use an iterative procedure in which an i th estimate of $\hat{\Gamma}_s(\omega)$ is used to obtain an $i+1$ st filter estimate. Some methods for modeling speech in this iterative framework are all pole modeling, the estimate maximum (EM) theory, etc.

Once the filter response $H(\omega)$ is found, the estimated spectrum of the speech signal will be

$$\hat{S}(\omega) = H(\omega)X(\omega) \quad (2.10)$$

It should be noted that $H(\omega)$ is a zero phase filter, that phase of original degraded speech $\angle X(\omega)$ is combined with $\hat{S}(\omega)$ to reconstruct the estimated speech signal $\hat{s}(n)$.

2.3 Adaptive Noise Canceling

Adaptive Noise Canceling refers to a class of adaptive signal processing algorithms that is based on the availability of a primary input source and a secondary reference source. The primary input source is assumed to speech plus additive noise as given by the relation in Eq. 2.2. The ANC filter acts on the reference channel and makes an estimation of the noise, which is then subtracted from the primary channel as shown in Fig. 2.2. The overall output of the canceler is used to control any adjustments made to the coefficients of the adaptive filter. The criterion for adjustment of these coefficients is generally to minimize the mean square energy.

2.4 Fundamental Frequency Tracking

This type of speech enhancement techniques depend on the tracking of fundamental frequency contour. This techniques depends on the underwritten fact that voiced speech is periodic, for which the frequency spectrum will be a line.

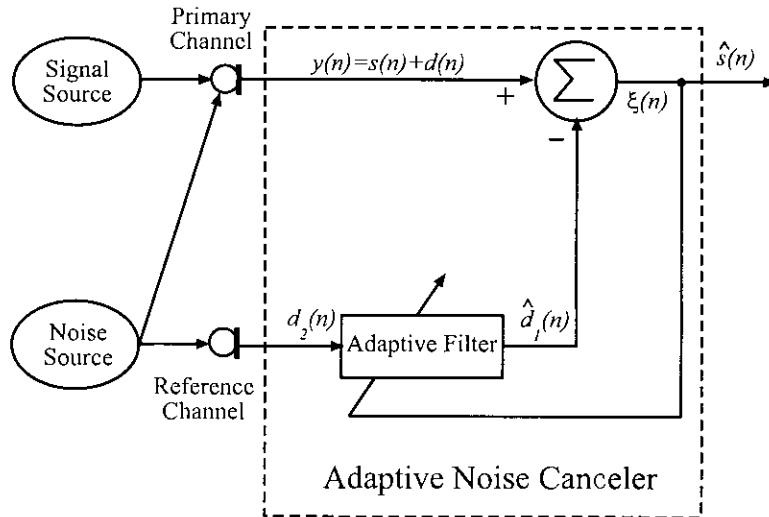


Figure 2.2: Flow diagram of ANC

Any spectral components between these lines represent noise that can be reduced. Some approaches for this fundamental frequency tracking are single channel ANC, Comb filtering etc.

2.5 Other Techniques

As an alternative to these traditional techniques and to conventional frequency domain speech processing theory, interest has emerged into studying speech as a nonlinear, dynamical system [38], [39]. Nonlinear time series methods perform analysis and processing in a reconstructed phase space, a time-domain vector space whose dimensions are time-lagged versions of the original time series [40]. The reconstructed phase space is therefore simply a plot of the time-lagged signal vectors, a parametric graph of the time series in which geometric structures of the underlying signal, called attractors or trajectories, appear. Reconstructed phase spaces have been shown to be topologically equivalent to the original system, if the embedding dimension is large enough [41]. This implies that the full dynamics of the system are accessible in this space, and for that reason, a phase space reconstruction potentially contains more information than a spectral representation [40], [42].

Chapter 3

Speech Enhancement using Perturbation

3.1 Introduction

Conventional speech enhancement algorithms, such those described in previous chapter can be broadly subdivided into two types. One is spectral subtraction and the other is Wiener Filtering. Both of them use an attenuation filter based on the assumed additive nature of the noise, that is interference between the speech signal and noise is constructive. As both noise and speech are stochastic process, there is an equal likelihood that noise is added to or subtracted from the clean speech signal. If both speech and noise are in same phase in FT domain (same sign for DCT, i.e. signal and noise both are either positive or negative), the magnitude of the noisy signal increases, magnitude decreases otherwise. It is observed that, for speech degraded with White Gaussian Noise (WGN), the number of constructive interference is much more than 50%. In fact, in pause periods this figure is almost 100% because of absence of any speech component. Perhaps this is the main reason behind the wide success of the two above mentioned filtration techniques.

As the interference between speech and noise may be destructive as well, an attenuating filtration on this noisy component will just lead to further distortion. Knowledge of the polarity of noise (i.e. whether constructive or destructive) would give rise to a new filtration technique where spectral subtraction methods would take the polarity of the noise coefficient into consideration. In this thesis it has been shown that the polarity of the noise coefficients can be identified to a great extent from noisy signal only.

I.Y. Soon and S.N. Koh investigated the polarity of noise when only the noisy signal is present [32]. They have considered the short-term stationary nature of speech and assumed that the clean speech coefficient remains reasonably unchanged. If speech is assumed to be stationary over some neighboring frames (say M frames, where M is odd), it can be assumed that the clean speech coefficient will be constant. If the magnitude of the coefficient in the current frame is lower than the mean of the magnitude of neighboring frames, it will be assumed that the interference of the noise is destructive and vice-versa. Though they have claimed the superiority of their algorithm over minimum mean square error (MMSE) filter by [9], they have not published the accuracy of their polarity estimation algorithm in that paper. However, the sign estimation part of their algorithm was reproduced and compared with actual sign. It was found that accuracy never exceeded 60%.

In this thesis a new technique is proposed for speech enhancement. A small perturbation is applied proportional to noise strength present in the noisy signal in DCT domain and its effect is observed in time domain to have an estimation of noise polarity and distribution as well. This estimated noise is then used to enhance the signal. Below, the step by step procedure of this technique is described:

3.2 Amount of Perturbation in Frequency Domain

Let $x_f(n)$ be a frame of noisy signal observation and $X_f(k)$ be its DCT. Similarly $s(n)$ and $v(n)$ be the signal and noise of that noisy frame, whereas $S(k)$ and $V(k)$ be their DCT respectively, such that:

$$x_f(n) = s(n) + v(n) \quad (3.1)$$

$$X_f(k) = S(k) + V(k) \quad (3.2)$$

To determine the amount of perturbation first an approximation of pure signal content and also the content of noise in the noisy signal ($X_f(k)$) are needed to be determined. To do this the average trend-line of the absolute value of $X_f(k)$ is determined by averaging $X_f(k)$ using small sub-bands. To reduce the abrupt changes in average values, these sub-bands are overlapped. A smooth

trend-line is obtained by fitting these average values using spline technique, let $\bar{X}_f(k)$ be the trend-line of $|X_f(k)|$.

As clean speech is band limited, it is likely that clean speech will not occupy the whole spectrum, rather a part of it (Fig. 3.1), whereas it is a well known fact that the frequency spectrum of white noise is random. If an average trend-line is determined, it will be almost flat (Fig. 3.2). This situation concludes that a sub-band with lowest energy in the average trend-line of noisy spectrum ($\bar{X}_f(k)$) will indicate the noise strength in that frame of noisy signal.

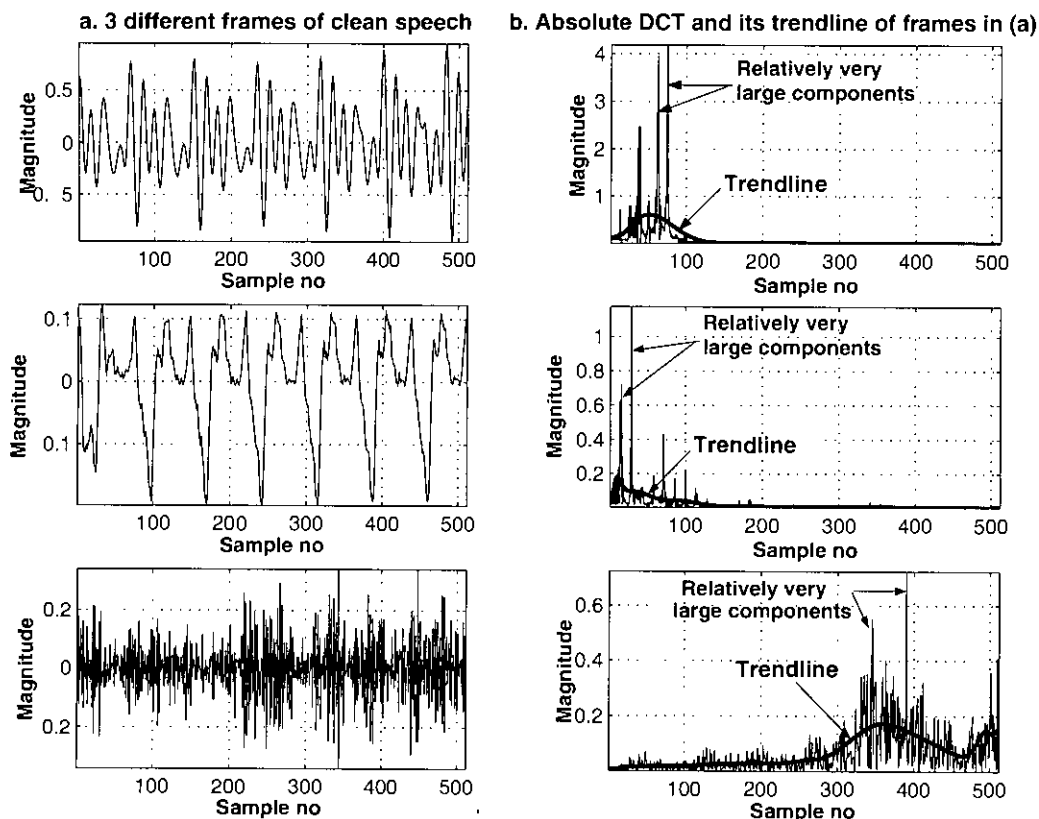


Figure 3.1: Three sample frames of clean speech and its DCT

Based on the above conclusion a floor is determined to have an approximation of average noise strength, by taking the average power of the lowest energy sub-band in DCT domain. Let it be $\bar{V}(k)$, as also shown in Fig. 3.3. Then approximation of clean speech strength $\bar{S}(k)$ (Fig. 3.3) can be found by subtracting noise content ($\bar{V}(k)$) from $\bar{X}_f(k)$ while taking zero ('0') for negative values, i.e.:

$$\bar{S}(k) = \max\{\bar{X}_f(k) - \bar{V}(k), 0\} \quad (3.3)$$

The strength of clean speech and that of noise so determined is then utilized to determine the amount of perturbation to be given in DCT domain.

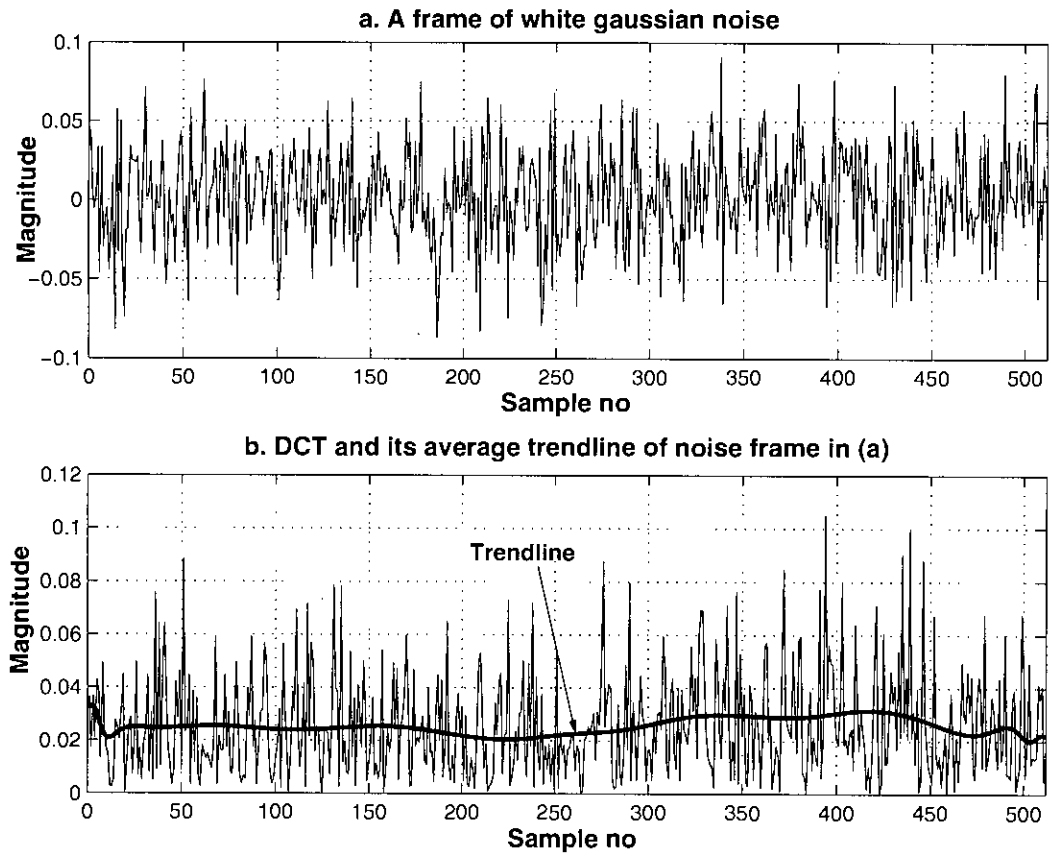


Figure 3.2: A frame of white gaussian noise and its DCT .

As the perturbation is targeted to reduce the noise only keeping signal content intact, the amount of perturbation applied, is proportional to the estimated average strength of noise present in noisy signal, i.e. ratio between $\bar{V}(k)$ and $[\bar{S}(k) + \bar{V}(k)]$. But for some abnormally high components (e.g., components around frequency index 50 in Fig. 3.3) $\bar{S}(k)$ is much smaller compared to actual signal present in $X_f(k)$. For those cases the ratio $\frac{\bar{V}(k)}{\bar{S}(k) + \bar{V}(k)}$ will result an over perturbation of the noise, leading to an over estimation of noise in time domain. Hence a threshold is used while applying perturbation. If the magnitude of noisy signal components $X_f(k)$ is greater than or equal to the twice as that of average noise strength $\bar{V}(k)$, then perturbation is proportional to the ratio between $\bar{V}(k)$ and $|X_f(k)|$. So the perturbed signal would be:

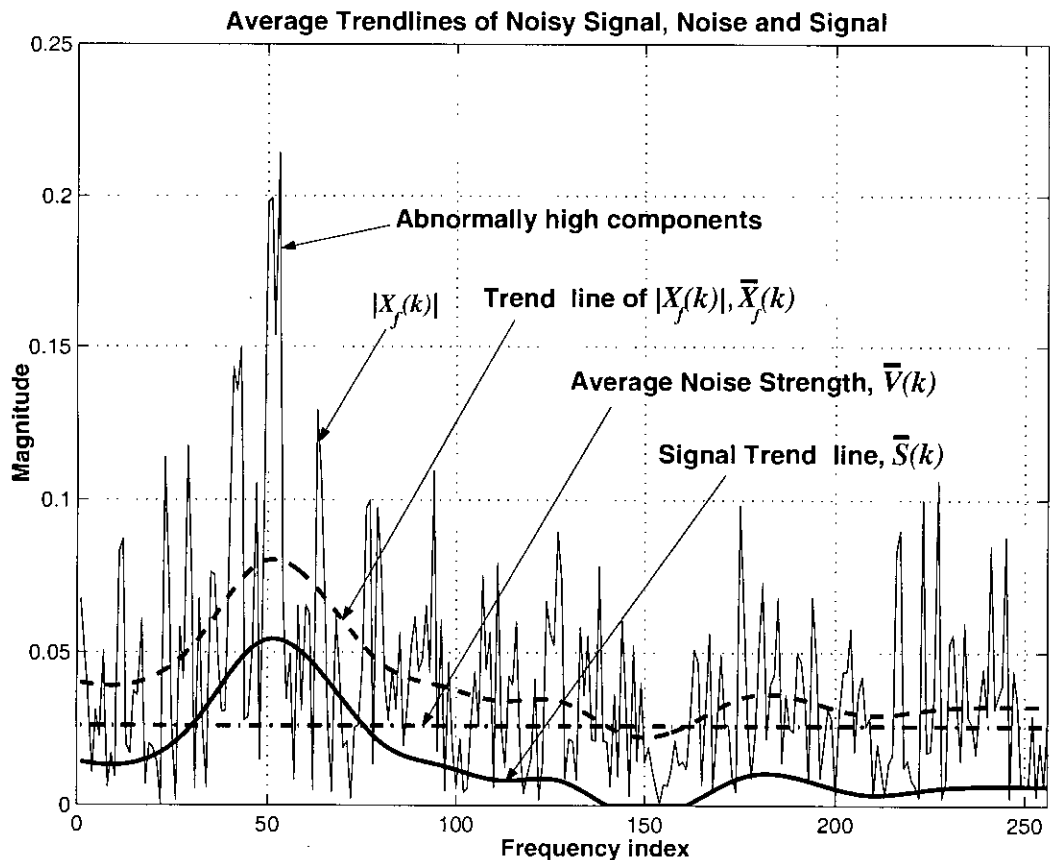


Figure 3.3: Sample trend-lines of a frame

$$\begin{aligned}
 X_p(k) &= X_f(k) \times \left\{ 1 - p \times \frac{\bar{V}(k)}{|X_f(k)|} \right\} && \text{If } X_f(k) \geq 2 \times \bar{V}(k) \\
 &= X_f(k) \times \left\{ 1 - p \times \frac{\bar{V}(k)}{\bar{V}(k) + \bar{S}(k)} \right\} && \text{Otherwise}
 \end{aligned} \tag{3.4}$$

Where p is the amount of perturbation, which is a very small number.

3.3 Estimation of Noise

The perturbed signal in time domain can be found using inverse DCT (IDCT) over $X_p(k)$ (Eq. 3.4), let it be $x_p(n)$. As the perturbation is given to reduce the effect of noise it is likely that the difference between $x_f(n)$ and $x_p(n)$ (i.e. $\{x_f(n) - x_p(n)\}$) will be dominated by the noise, which can be taken as an unscaled estimation of noise, though this will contain a rather small part of clean signal. Let this unscaled estimation of noise be $v'_e(n)$ which can be given

as:

$$\begin{aligned}
 v_e'(n) &= p \times v(n) + q \times s(n) \\
 &= p \left\{ v(n) + \frac{q}{p} \times s(n) \right\} \\
 &= p \{ v(n) + \beta \times s(n) \}
 \end{aligned} \tag{3.5}$$

As already mentioned p is the amount of perturbation and q is the residual signal present in the difference signal, which is very small according to our assumption. And β is the ratio between q and p . Then error function (let ξ) between actual noise and its estimation is given by:

$$\begin{aligned}
 \xi &= v(n) - \alpha v_e'(n) \\
 &= v(n) - \alpha p \{ v(n) + \beta s(n) \}
 \end{aligned} \tag{3.6}$$

Where α is a scaling factor such to minimize the sum of square of error function ξ , i.e. $\sum |\xi|^2$. Now to minimize $\sum |\xi|^2$ we have

$$\begin{aligned}
 \frac{d \sum |\xi|^2}{d\alpha} &= 0 \\
 \frac{d}{d\alpha} \sum \{ v(n) - \alpha v_e'(n) \}^2 &= 0 \\
 \frac{d}{d\alpha} \left[\sum \{ v^2(n) - 2\alpha \times v(n) \times v_e'(n) + v_e'^2(n) \} \right] &= 0 \\
 -2 \sum v(n) \times v_e'(n) + 2\alpha \sum v_e'^2(n) &= 0
 \end{aligned}$$

Then α will be

$$\begin{aligned}
 \alpha &= \frac{\sum v(n)v_e'(n)}{\sum v_e'^2(n)} \\
 &= \frac{\sum p \times v(n) \{ v(n) + \beta s(n) \}}{\sum p^2 \{ v(n) + \beta s(n) \}^2} \\
 &= \frac{p \left[\sum v^2(n) + \beta \sum v(n) \times s(n) \right]}{p^2 \left[\sum v^2(n) + 2\beta \sum v(n) \times s(n) + \beta^2 \sum s^2(n) \right]} \\
 &= \frac{1}{p} \frac{\sum v^2(n)}{\sum v^2(n) + \beta^2 \sum s^2(n)} \quad \left[\because \sum v(n) \times s(n) = 0 \right] \\
 &= \frac{1}{p} \frac{1}{1 + \beta^2 \frac{\sum s^2(n)}{\sum v^2(n)}} \quad \text{As } s(n) \text{ and } v(n) \text{ are uncorrelated.} \\
 &= \frac{1}{p} \frac{1}{1 + \beta^2 \frac{\sum s^2(n)}{N \times P(v)}}
 \end{aligned}$$

$$\therefore \alpha = \frac{1}{p} \frac{1}{1 + \beta^2 \frac{\sum s^2(n)}{N \times P(v)}} \quad (3.7)$$

Where N is the frame size and $P(v)$ is the average noise power of that particular frame. Then the estimated noise will be

$$\hat{v}_e(n) = \alpha \times v'_e(n) \quad (3.8)$$

And enhanced signal will be:

$$\hat{s}(n) = x_f(n) - \hat{v}_e(n) \quad (3.9)$$

3.4 Estimation of Noise Power ($P(v)$)

This is the most critical part in this algorithm. Because, the performance of the algorithm depends on the accuracy of the estimation of the noise power of the frame. In this work two methods have been employed for this purpose. Both methods have their own advantages and disadvantages.

In the first method we have used a global noise power, which is calculated globally from pause interval. A number of algorithms is present to determine the pause interval [43], [45], [46]. This method gives better result in term of performance and quality, particularly for stationary noise.

In the second method the noise power level is estimated taking into account that the speech component is not dominant at the finest level. It is known that the kurtosis for WGN is 3 [47]. 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 identify the frequency band dominated by noise, i.e., noise distribution in a particular frame. The noisy signal frame is subdivided into sub-frames and kurtosis of those sub-frames are analyzed for the identification of noise. If the value of kurtosis is less than or equal to 3 then it is assumed that, that sub-frame is a noise dominated one.

3.5 Conclusion

Speech enhancement using a new perturbation technique has been presented in this chapter. Main focus of this chapter is to develop a mathematical base to apply the perturbation techniques to enhance the quality of speech. The performance of the algorithm, i.e. quality and intelligibility of the enhanced speech will be discussed in the next chapter.

Chapter 4

Simulation Results

4.1 Data Used for Simulation

The proposed enhancement algorithm is tested for a data set consisting of 40 different utterances from the TIMIT speech database. Half of the utterances are from male speakers and half are from female speakers. 8 kHz is chosen as the sampling rate for speech signals, because of its wide use in communication channels. Also 16 bits are used to encode each of the speech samples.

The noise type for simulation is chosen to be computer generated WGN. WGN is chosen for its wide use in studying the performance of enhancement systems and also it is a good model for wide-band noise sources which are often encountered in practice, e.g., thermal noise in communication systems. Each speech signal is degraded by WGN with overall SNR in the range from -5 to 30 dB.

Individual and aggregated results are presented here. The simulation results are compared with three existing methods. First two are the modified power spectral estimator and the modified parametric spectral subtraction method proposed by Hasan *et al.* in [35], which is referred as MPE and PARA, respectively, in this thesis. These two techniques are quite recent and show very good enhancement result. Third one is the minimum mean square error (MMSE) log-spectral amplitude estimator proposed by Ephraim *et al.* in [36], which is referred as MMSE-LSAE in this thesis. This is one of the fundamental works in speech enhancement. The proposed two techniques as mentioned in section 3.4 is referred as ‘Pert1’ for first method and ‘Pert2’ for second method.

4.2 Simulation Parameters

The frame size for the simulation is chosen to be 32mS, equivalent 256 samples with an overlap of 128 samples. For determining average trend-line and also for finding noise floor a sub-band of 32 samples is used with an overlap of 16 samples. No window function has been used. The perturbation factor p is taken as 0.01(1%).

As shown in Eq. 3.7 (presented below for convenience) that, to determine

$$\alpha = \frac{1}{p} \frac{1}{1 + \beta^2 \frac{\sum s^2(n)}{N \times P(v)}}$$

α , a constant β is used, which is to be determined empirically. In fact, the assumption that speech and noise are uncorrelated (i.e. $\sum v(n) \times s(n) = 0$) is not entirely true for such a small time frame of 32mS. This led to a conclusion that β should not be constant, rather be dependant on SNR. In the simulation it is found that β not only depends on overall SNR, but also on specific frame-based SNR. For a particular overall SNR, SNR for different speech frames may be different. For example SNR in pause period is very low, while SNR in speech dominant region is very high.

In simulation a set of β value is chosen empirically. As mentioned in section 3.3, that β indicates the residual signal (Eq. 3.5) present in unscaled estimation of noise. For high SNRs the noise is very small compared to signal, as a result any perturbation applied will contain a higher proportion of signal, i.e. such a perturbation will affect the signal more than it does in case of low SNRs. So for high SNRs the residual signal will be more pronounced, indicating that β should be made higher for high SNRs. Also correct identification of noise polarity influences the choice of β . If there is no residual signal present (i.e. $\beta = 0$), then α would be $\frac{1}{p}$ (equal to 100 for our case). But the fact that polarity identification is better for low SNRs, rather than that of high SNRs (will be shown in next section 4.3), lead to a decision that β should be higher for higher SNR, such that the value of α is lowered to minimize the distortion introduced by wrong polarity identification.

The table for β is given in Table 4.1. A set of values is chosen an overall input SNR, and from that set an individual value is chosen for an individual frame-based SNR. For example, for an estimated overall SNR of 5dB and a frame-based SNR of 10dB, value in 5dB column and 10dB row is used (0.04 in

this case). The table is given for an overall SNR range of -5dB to 30dB , because noise with SNR below -5dB is practically unrealistic and that with over 30dB is practically treated as noise free.

Table 4.1: β values chosen for various SNR

Frame SNR	Overall SNR							
	-5	0	5	10	15	20	25	30
-30	0.0770	0.0770	0.0770	0.0770	0.0770	0.0770	0.0770	0.0770
-25	0.0742	0.0742	0.0742	0.0742	0.0742	0.0742	0.0742	0.0742
-20	0.0714	0.0714	0.0714	0.0714	0.0714	0.0714	0.0714	0.0714
-15	0.0686	0.0686	0.0686	0.0686	0.0686	0.0686	0.0686	0.0686
-10	0.0658	0.0658	0.0658	0.0658	0.0658	0.0658	0.0658	0.0658
-5	0.0630	0.0630	0.0630	0.0630	0.0630	0.0630	0.0630	0.0630
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0010	0.0100	0.0400	0.1000	0.1000	0.2000	0.2500	0.3000
10	0.0010	0.0100	0.0400	0.1000	0.1000	0.2000	0.2500	0.4400
15	0.0010	0.0100	0.0400	0.1000	0.1000	0.2000	0.2500	0.5800
20	0.0010	0.0100	0.0400	0.1000	0.1000	0.2000	0.2500	0.7200
25	0.0010	0.0100	0.0400	0.1000	0.1000	0.2000	0.2500	0.8600
30	0.0010	0.0100	0.0400	0.1000	0.1000	0.2000	0.2500	1.0000

Although the value of α is kept constant for a frame (as per Eq. 3.7), it is observed that, using α as a vector (say α') determined by the following Eq. 4.1 produce better results in simulation.

$$\alpha'(n) = \frac{\gamma}{p} \frac{1}{1 + \frac{\beta^2 \times s^2(n)}{P(v)}} \quad (4.1)$$

Here γ is a multiplying factor that lies within unity. This factor is introduced to incorporate the dependency of α on overall SNR.

To determine the characteristic of the factor γ in Eq. 4.1, simulation is done (averaged result for both male and female voice with 20 different utterances from 20 individual runs) sweeping γ from 0.7 to 1.5 with an increment of 0.05. The overall SNR improvement is observed and result is given in Fig. 4.1. The locus of peaks for different input SNRs is also plotted. The two locus (male and female) is averaged and best fitted for 1st order using least square technique that gives a characteristic equation for γ , which is given in Eq. 4.2:

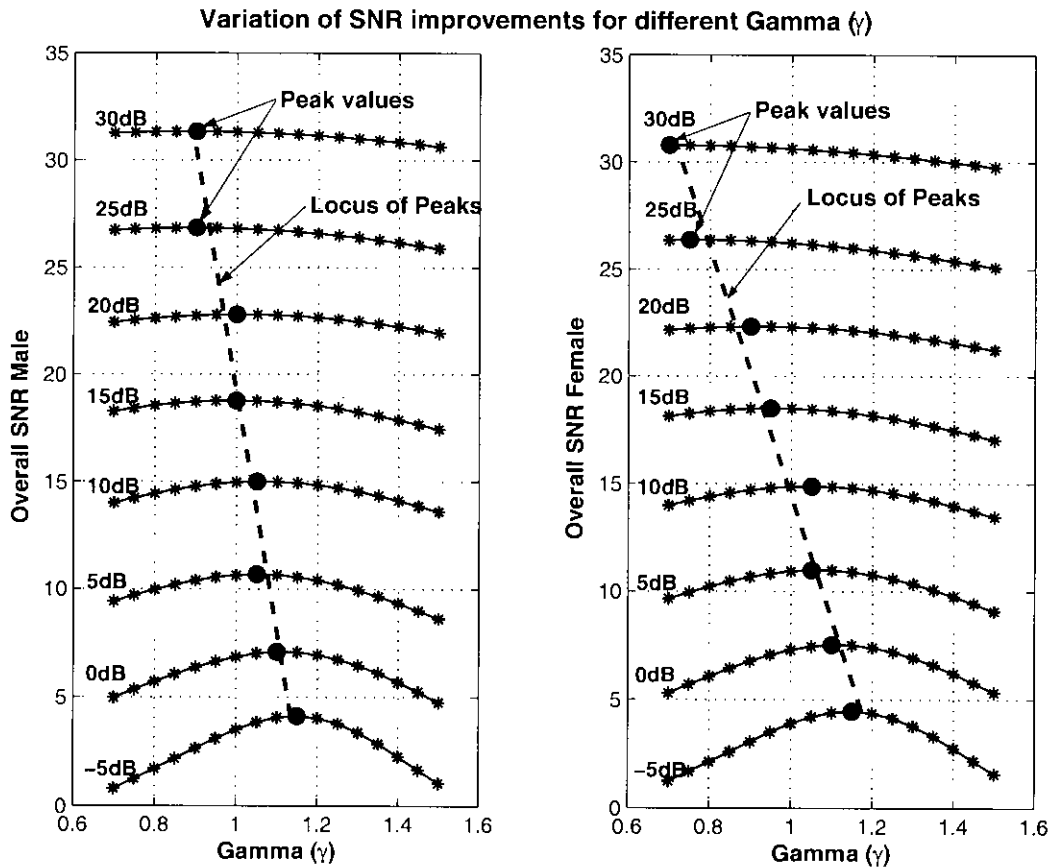


Figure 4.1: Characteristics of the factor Gamma (γ)

$$\gamma = -0.01 \times SNR_{input} + 1.1125 \quad (4.2)$$

4.3 Identification of Noise Polarity

As mentioned in section 3.1, most of the researchers have ignored the noise polarity with respect to clean signal. I. Soon *et al.* investigated the the polarity of noise in [32]. However he did not mention any numerical values of correct identification. The part of the algorithm that identifies polarity is implemented and its result is presented in Table 4.2, referred as SOON in column 2 and 5, along with results of noise polarity identification for the proposed two methods. The result presented in Table 4.2 is the average result of 20 individual runs of 20 different utterances, both for male and female.

If the difference of magnitude of noisy signal and perturbed signal is positive then the noise is additive and vice-versa. To calculate the percentage of correct

Table 4.2: Noise polarity identification

Input SNR	% Identified correctly for Male Speaker			% Identified correctly for Female Speaker		
	SOON [32]	Pert1	Pert2	SOON [32]	Pert1	Pert2
-5	53.393	89.645	89.645	53.019	89.148	89.148
0	55.038	87.884	87.884	55.022	87.077	87.077
5	56.184	86.101	86.100	56.880	85.052	85.051
10	56.605	83.775	83.770	58.235	82.499	82.495
15	56.351	81.335	81.329	58.817	79.519	79.510
20	55.605	78.291	78.282	58.512	75.984	75.935
25	54.657	74.644	74.632	57.360	72.243	72.196
30	53.872	70.827	70.724	55.810	68.354	68.211

identification, the calculated polarity is compared with that of actual polarity, that is the difference of magnitude of noisy signal and clean signal, being additive if the difference is positive, and subtractive otherwise. The results are presented graphically in Fig. 4.2. From results it is evident that correct polarity identification by Soon is around 50%, which is mediocre. Whereas in proposed method this rate is rather high, being around 90% for lower SNRs. In high SNR the signal amplitude is affected even for a small perturbation as the magnitude of noise is very low compared to that of signal, that results lower percentage of correct identification. Still the value is around 70%. Also it is evident that the identification rate is higher for male speaker. This is because the female voice is more extended in frequency spectrum than that of male, for which, the perturbation, which is targeted for noise only, affects signal magnitude also.

4.4 Performance Test

The assessment of the proposed method is based on objective and subjective quality tests. For objective tests classical overall SNR, Average Segmental SNR and Itakura-Saito (IS) distortion measure are used [5]. For subjective tests speech spectrograms and informal listening are used.

Enhancement results of a male utterance and that of a female utterance is shown in Fig. 4.3 and Fig. 4.4 respectively. The denoised signal enhanced by MPE, PARA and MMSE-LSAE, along with that with the proposed methods

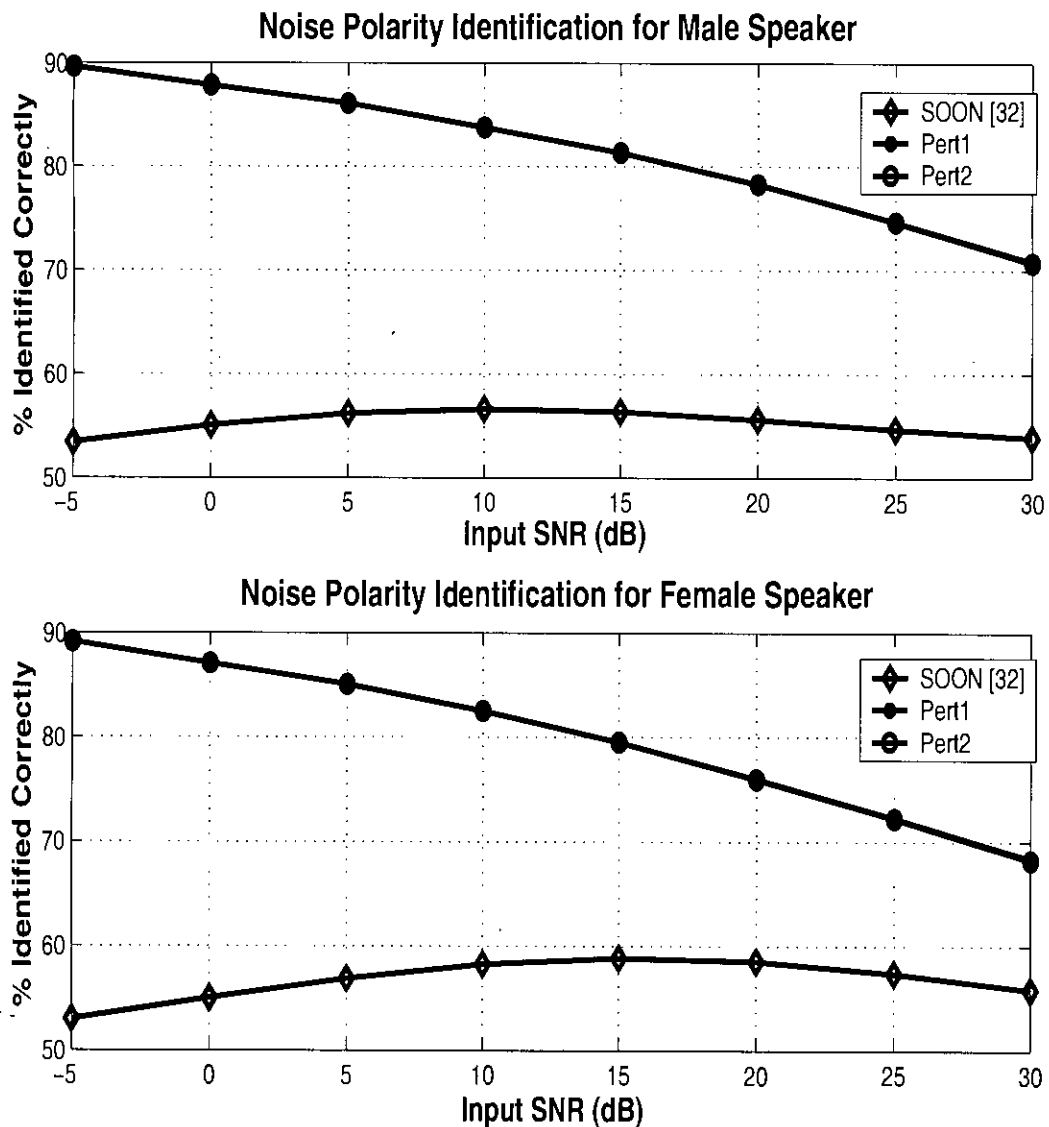


Figure 4.2: Polarity identification for different methods

(Pert1 and Pert2) are presented in these figures. In both cases clean signal and noisy signal are presented for comparison. It is clear from the figures that the proposed methods enhance better in pause periods. This is because in frames locating around pause interval, the SNR is very low, and the proposed algorithm identifies noise better for such low SNRs, which is already mentioned in section 4.3.

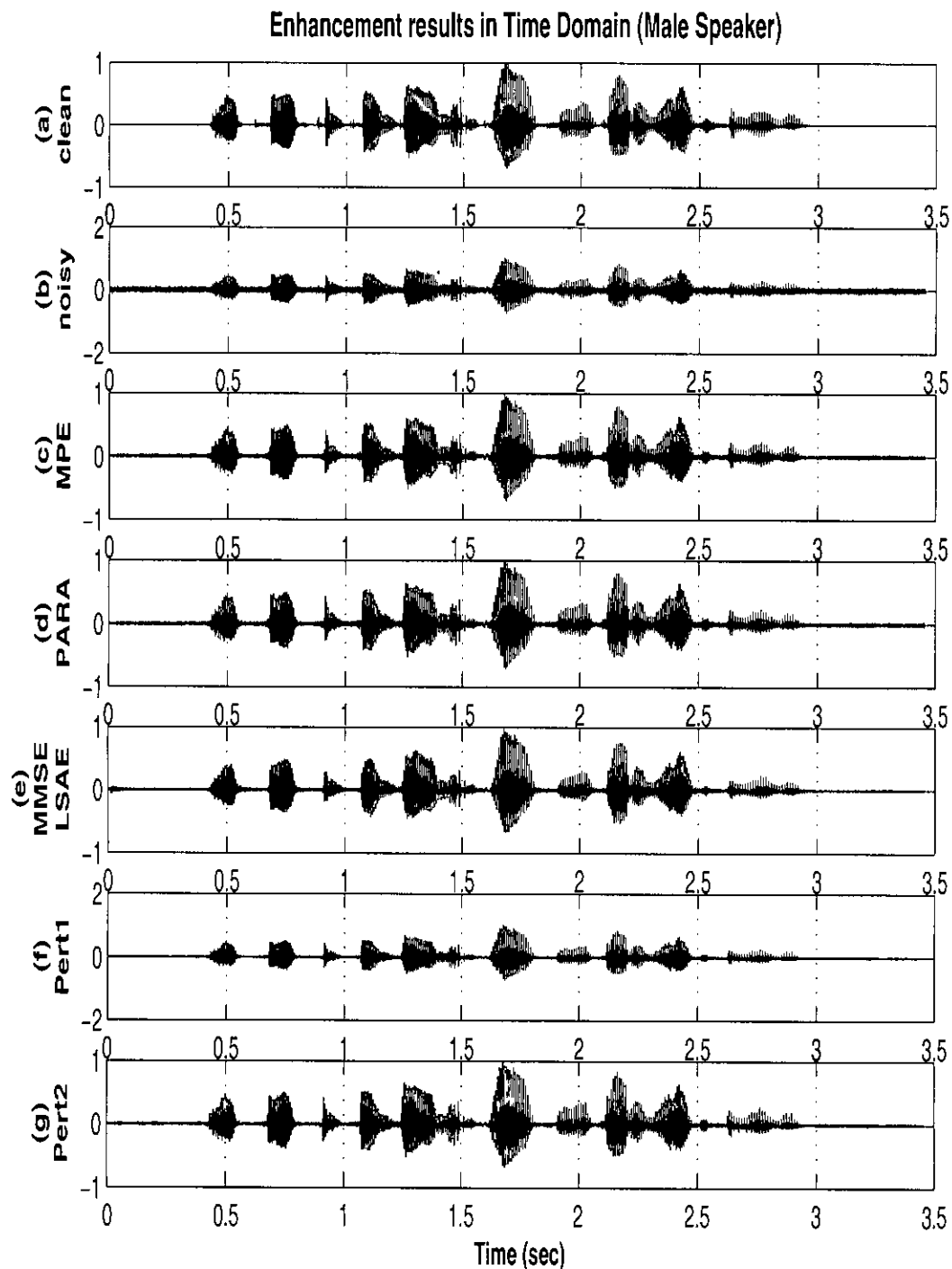


Figure 4.3: Enhancement results for male utterance “Would you please confirm the government policy regarding waste removal”: (a) clean; (b) noisy (corrupted by 10dB WGN); enhanced using (c) MPE; (d) PARA; (e) MMSE-LSAE; (f) proposed Pert1 and (g) Proposed Pert2.

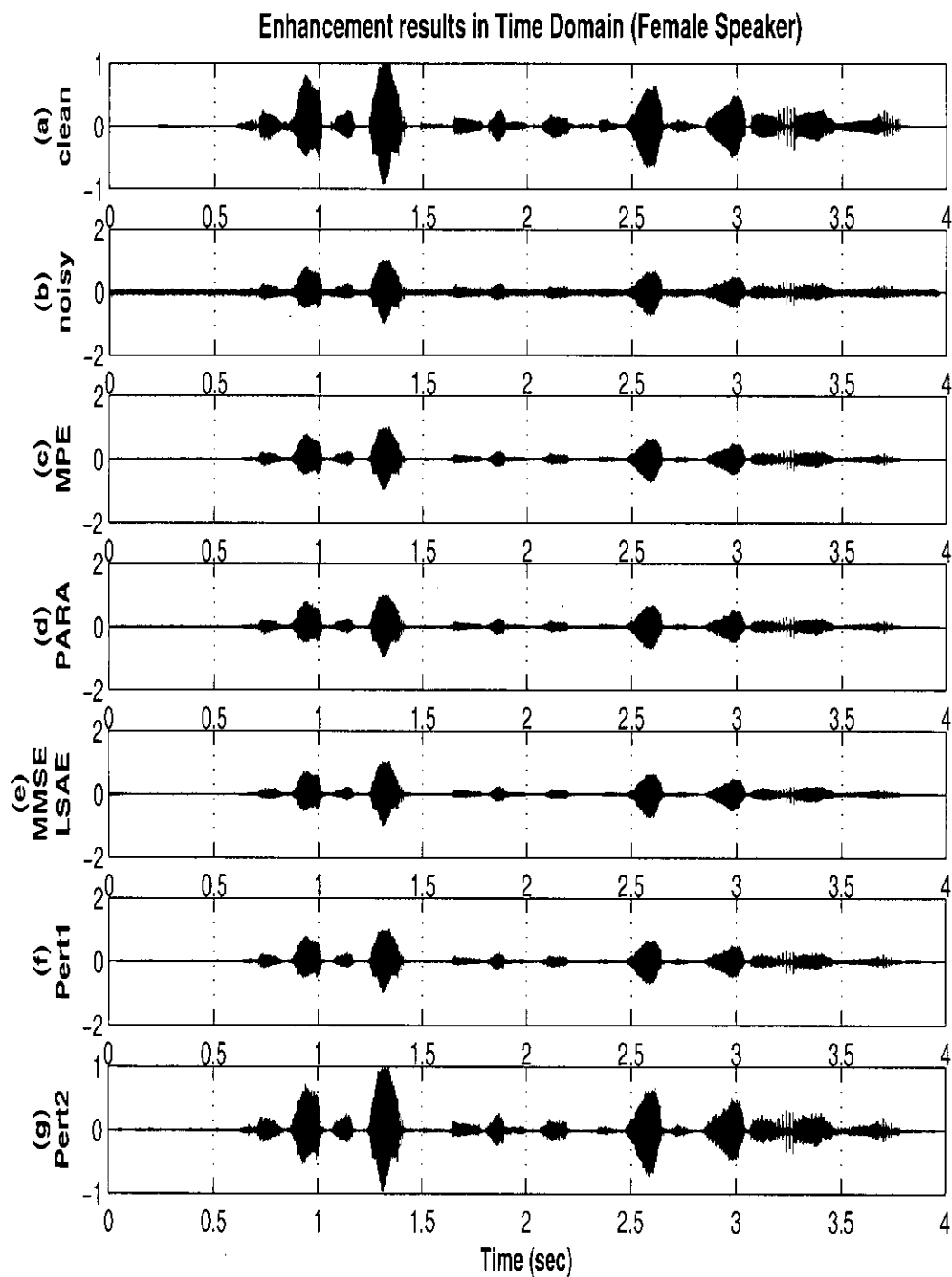


Figure 4.4: Enhancement results for female utterance “She had your dark suit in greasy wash water all year”: (a) clean; (b) noisy (corrupted by 10dB WGN); enhanced using (c) MPE; (d) PARA; (e) MMSE-LSAE; (f) proposed Pert1 and (g) Proposed Pert2.

4.5 Objective Tests

4.5.1 Overall SNR

The SNR is the most widely used measure for assessing enhancement algorithms for broadband noise distortions. The SNR of enhanced signal $\hat{s}(n)$ (in dB) is defined as:

$$\text{SNR} = 10 \times \log_{10} \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} [\hat{s}(n) - s(n)]^2} \quad (4.3)$$

The average results of 20 independent runs correspond to 20 different utterances, both for male and female speakers, are given in Table 4.3. In the 3rd, 4th and 5th columns, SNR of the improved speech is given for methods MPE, PARA and MMSE-LSAE respectively. Results that of the proposed two methods as described in section 3.4 is given in columns 6 and 7.

The resulting plot of Improved SNR vs. input SNR is given in Fig. 4.5 for male speaker and in Fig. 4.6 for female speaker. It is evident that for male speaker the proposed method ‘Pert1’ show better result for low and moderate SNRs. For higher SNRs (more than 20dB) MPE method is proved to be better. For high SNR magnitude of noise is negligible with respect to that of signal. Which implies that any perturbation on noise will affect signal more, which is undesirable. So estimation of noise is more erroneous for high SNRs. Also notice that improvement result for male speaker is better than the female ones. This is due to the fact that female voice is more extended in frequency domain than male voice as they contain more high frequency element. So noise at the finest level interact more with speech for female voice, which in turn affect the signal components while applying perturbation. Resulting a lower polarity identification and lower improvement. The proposed ‘Pert2’ results decline of overall SNR for high input SNRs. As mentioned in section 3.4 the noise power level is estimated using kurtosis with a threshold value 3.0. In fact kurtosis of some speech sub-band is below 3.0, which result over estimation of noise, adding additional distortion to the enhanced signal.

4.5.2 Average Segmental SNR

The overall SNR represents an average error over time and frequency for a processed signal. Since the correlation of SNR with subjective quality is so poor,

Table 4.3: Overall SNR improvement for various noise levels, obtained using MPE, PARA, MMSE-LSAE and proposed two methods

Speaker	Input SNR [dB]	Improved SNR for Methods				
		MPE	PARA	MMSE LSAE	Prop. Method	
					Pert1	Pert2
Male	-5	2.36	2.09	2.44	4.13	4.14
	0	6.54	6.26	5.86	7.10	7.10
	5	10.41	10.12	9.11	10.67	10.43
	10	14.45	14.01	12.63	14.96	14.28
	15	18.57	17.86	16.57	18.75	17.43
	20	22.79	21.56	20.78	22.76	20.92
	25	27.16	25.30	25.18	26.85	23.44
	30	31.70	28.88	29.53	31.32	25.75
Female	-5	3.67	3.33	4.36	4.46	4.47
	0	7.57	7.25	7.47	7.53	7.53
	5	11.28	10.94	10.57	10.98	10.87
	10	14.97	14.49	13.77	14.86	14.45
	15	18.76	18.09	17.18	18.50	17.57
	20	22.71	21.70	20.88	22.32	20.74
	25	26.85	25.19	24.92	26.34	23.35
	30	31.23	28.66	29.24	30.75	25.77

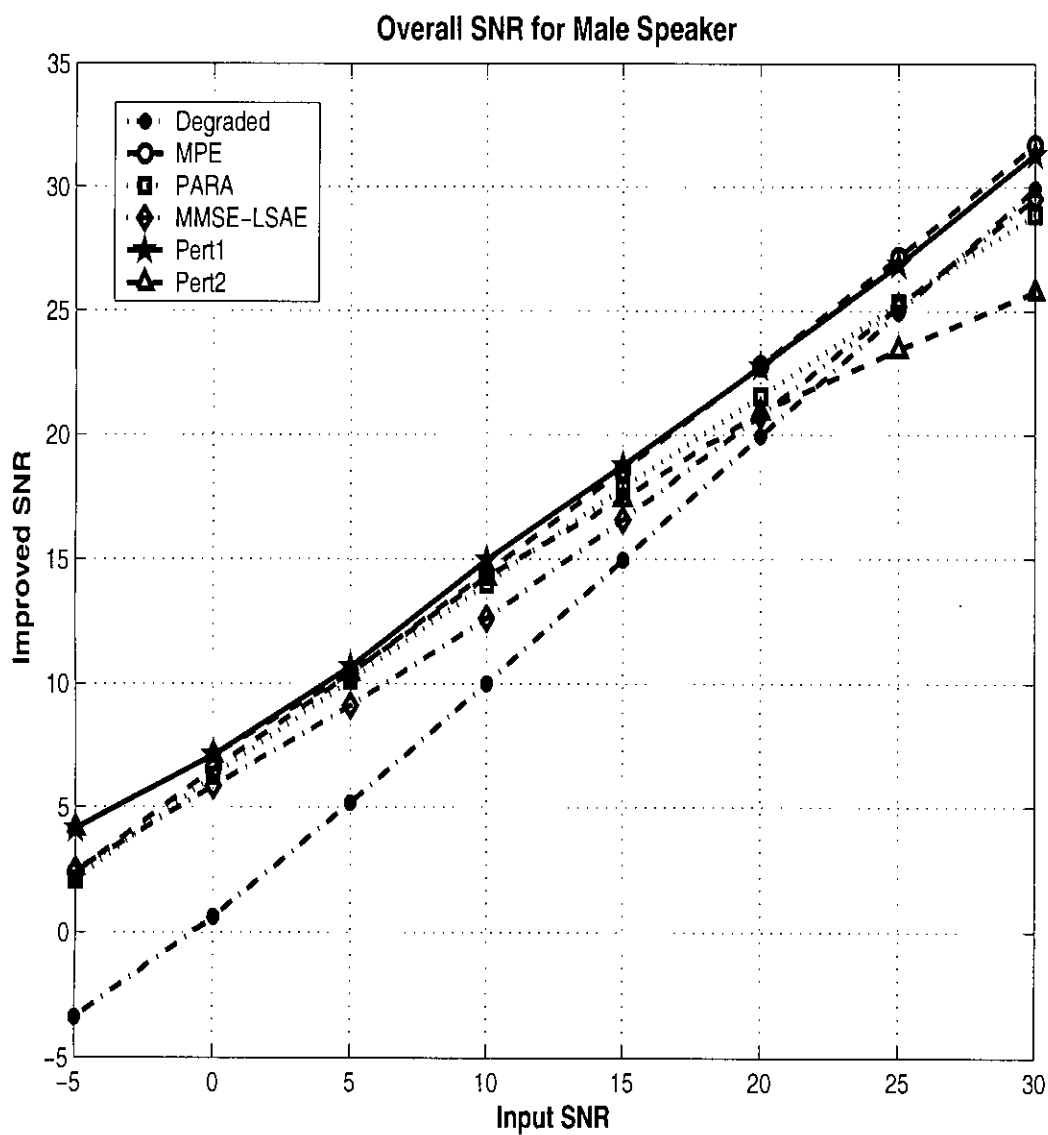


Figure 4.5: Overall SNR for male utterance: Result is averaged using results of 20 utterances

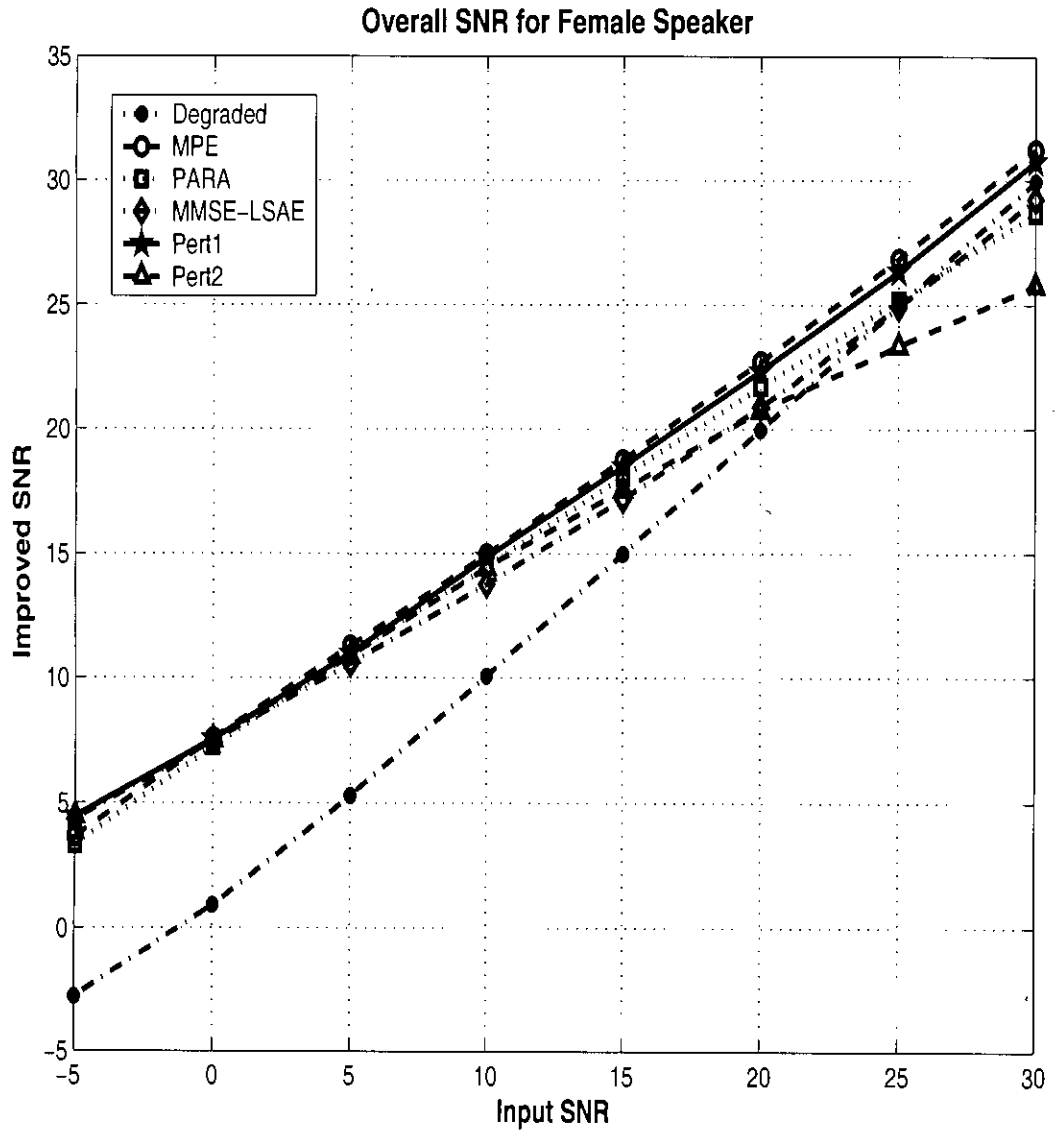


Figure 4.6: Overall SNR for female utterance: result is averaged using results of 20 utterances

it is of little interest as a general objective measure of speech quality [48]. Instead, the frame-based segmental SNR is chosen, which is a reasonable measure of speech quality. It is known as *segmental SNR* (SNR_{seg}), and is formulated by averaging frame level SNR estimates as follows:

$$SNR_{seg} = \frac{1}{M} \sum_{j=0}^{M-1} 10 \times \log_{10} \left[\frac{\sum_{n=m_j-N+1}^{m_j} s^2(n)}{\sum_{n=m_j-N+1}^{m_j} [\hat{s}(n) - s(n)]^2} \right] \quad (4.4)$$

Where m_0, m_1, \dots, m_{M-1} are the end times for the M frames, each of which is of length N . Frames with very high SNRs (above $30 \sim 35$ dB) do not reflect large perceptual differences. Likewise, during periods of silence, SNR values can become very negative since signal energies are small. These frames do not truly reflect the perceptual contributions of the signal. Therefore, thresholds are often set both for higher and lower SNR to provide a bound on frame based SNRs. Here lower threshold is set to $-10dB$, and that of higher is $35dB$ as in [5].

As in the case of overall SNR, the average results of 20 independent runs correspond to 20 different utterances, both for male and female speakers, are given in Table 4.4. Improvement results for MPE, PARA and MMSE-LSAE are given in column 3, 4 and 5. Results that of the proposed two methods as described in section 3.4 is given in columns 6 and 7.

The plot of Improved Segmental SNR against each input SNR, for male speaker, is given in Fig. 4.7 and that of female is given in Fig. 4.8. It is evident that proposed method 'Pert1' show better result than previous methods for almost all input SNRs. The proposed 'Pert2' results decline of overall SNR for high input SNRs. The same reasoning is applicable for improvement as mentioned in previous sub-section (4.5.1).

4.5.3 Itakura-Saito Distortion Measure

For an original clean frame of speech with linear prediction (LP) coefficient vector, \vec{a}_ϕ , and processed speech coefficient vector, \vec{a}_d , the Itakura-Saito distortion measure is given by,

$$d_{IS}(\vec{a}_d, \vec{a}_\phi) = \left[\frac{\sigma_\phi^2}{\sigma_d^2} \right] \left[\frac{\vec{a}_d R_\phi \vec{a}_d^T}{\vec{a}_\phi R_\phi \vec{a}_\phi^T} \right] + \log \left(\frac{\sigma_d^2}{\sigma_\phi^2} \right) - 1 \quad (4.5)$$

Where σ_d^2 and σ_ϕ^2 represent the all-pole gains for the processed and clean speech frame respectively and R_ϕ is the auto-correlation lags for clean speech.

Table 4.4: Segmental SNR improvement for various noise levels, obtained using MPE, PARA, MMSE-LSAE and proposed two methods

Speaker	Input SNR [dB]	Degraded SNR_{seg}	Improved SNR_{seg} for Methods				
			MPE	PARA	MMSE LSAE	Prop. Method	
						Pert1	Pert2
Male	-5	-6.52	-4.24	-4.36	-4.02	-2.97	-2.97
	0	-4.67	-1.97	-2.13	-1.96	-1.20	-1.20
	5	-2.30	0.44	0.24	0.12	0.96	0.85
	10	0.55	3.09	2.78	2.43	3.63	3.36
	15	3.76	5.95	5.45	5.31	6.30	5.79
	20	7.18	9.24	8.39	8.73	9.31	7.70
	25	10.79	13.06	11.80	12.40	12.74	11.62
	30	14.55	17.06	15.42	15.91	16.43	14.70
Female	-5	-6.43	-3.57	-3.74	-2.85	-2.54	-2.53
	0	-4.67	-1.19	-1.39	-0.59	-0.53	-0.53
	5	-2.27	1.45	1.19	1.78	1.79	1.75
	10	0.74	4.28	3.91	4.26	4.55	4.39
	15	4.26	7.31	6.82	6.90	7.35	7.03
	20	8.16	10.56	9.83	9.85	10.34	9.83
	25	12.30	14.15	12.98	13.23	13.74	12.70
	30	16.43	17.98	16.38	16.85	17.48	15.68

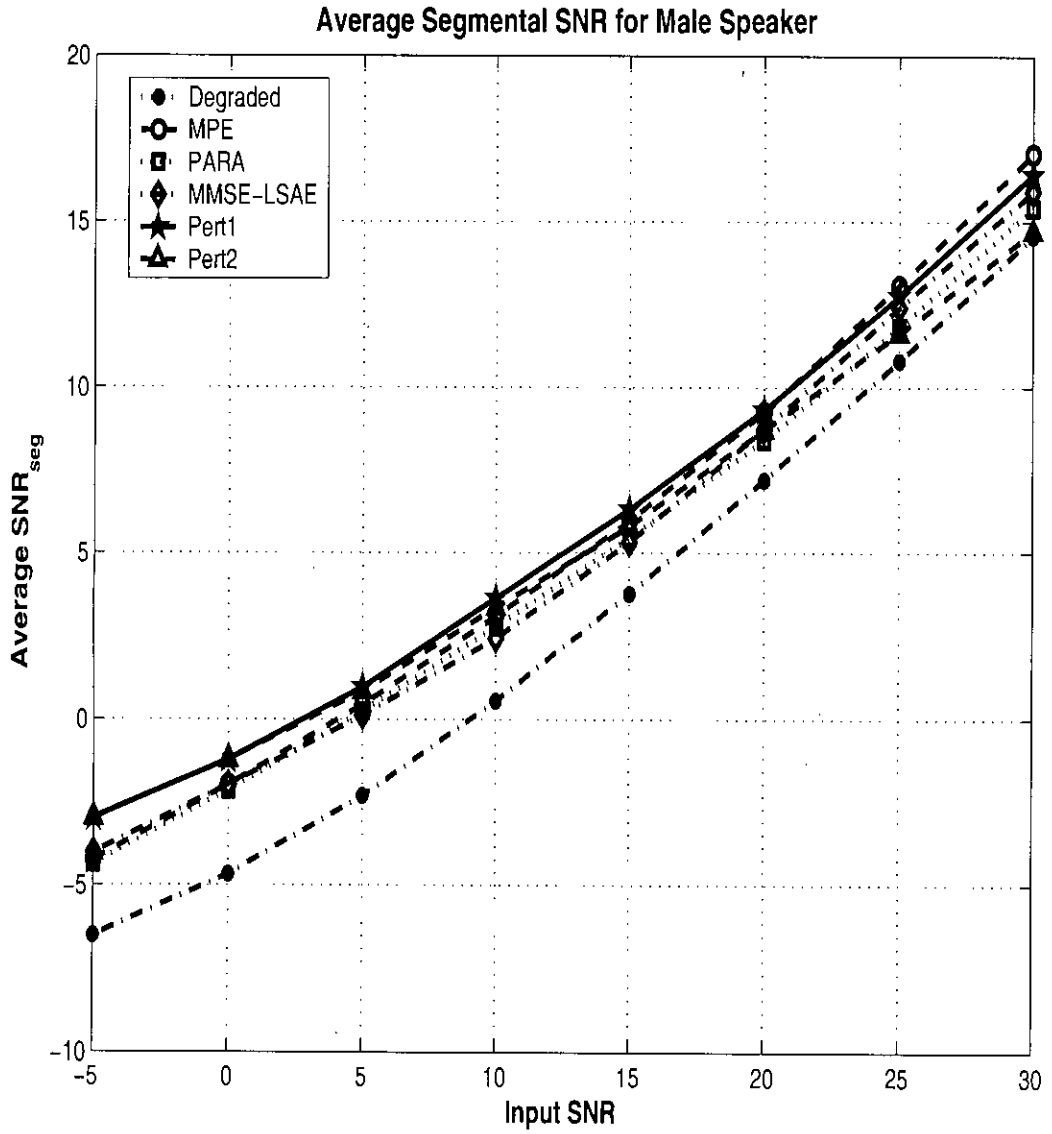


Figure 4.7: Average Segmental SNR improvement for male utterance: result is averaged using results of 20 utterances

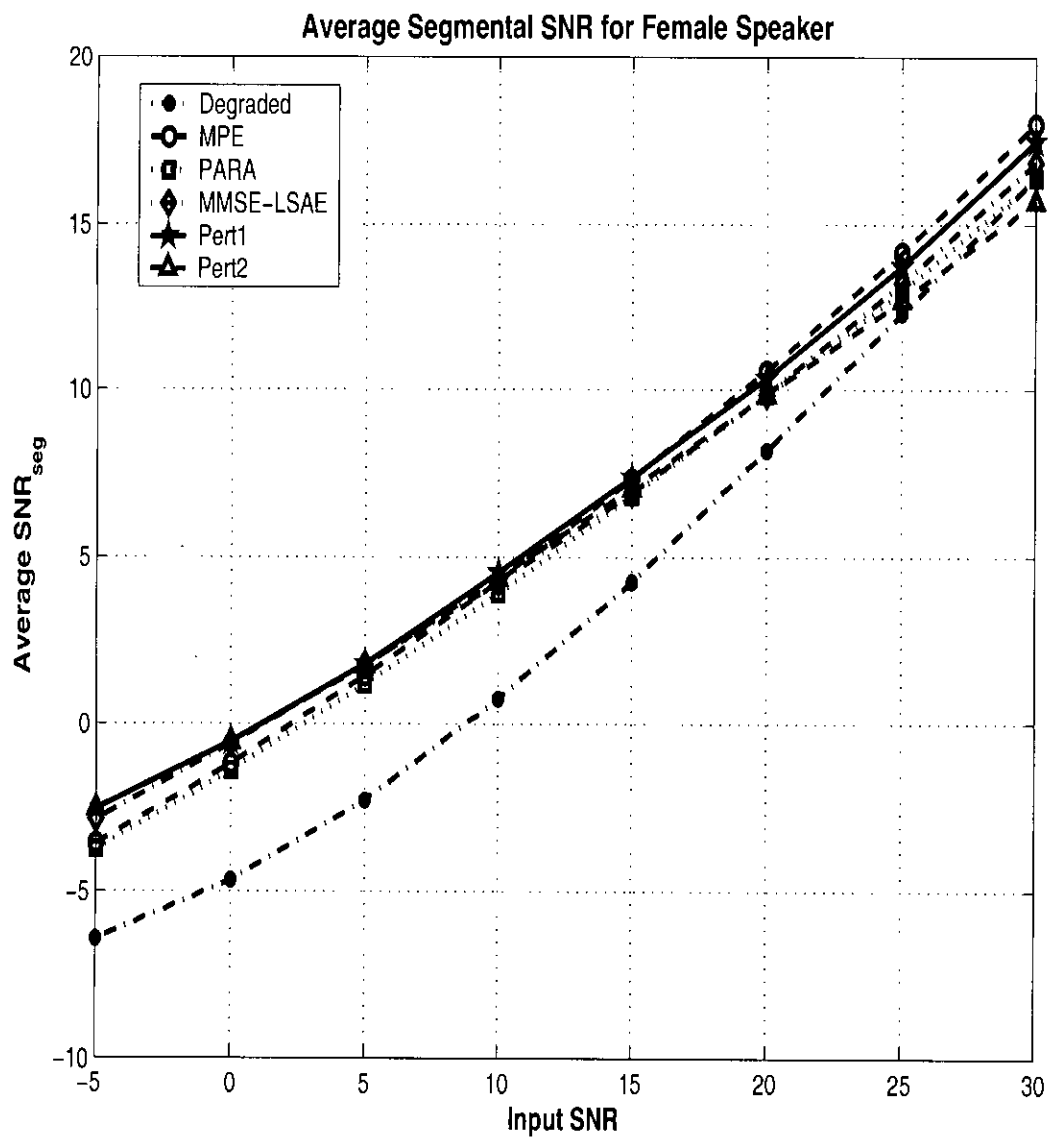


Figure 4.8: Average Segmental SNR improvement for female utterance: result is averaged using results of 20 utterances

There are several ways to obtain overall quality scores. For most measures, finding a mean across a large test set is reasonable. If users want a general measure of performance the median of the resulting frame-level scores is more useful (a mean quality measure is typically biased by a few frames in the tails of the quality measure distribution). Another way to get a reasonable overall measure is to find the mean using the first 95% of the frames. This allows for the removal of a fixed number of frames which may have unrealistically high distortion levels. Results are presented in Table 4.5, plot of which is given in Fig. 4.9 for male speaker and Fig. 4.10 for female speaker. Proposed method ‘Pert2’ result a very high amount of distortion, none of the methods ‘Pert1’ and ‘Pert2’ show any better result than that of MPE. For clean speech, energies of unvoiced segments are comparable to those of noise. Applying perturbation to those segments, it results an over-estimation of noise. Consequently those unvoiced segments reduced in magnitude, resulting abnormally high distortion at some points.

4.5.4 Spectrogram

Spectrogram is a time-dependent Fourier transform for a sequence. The time-dependent Fourier transform is the discrete-time Fourier transform for a sequence, computed using a sliding window. The spectrogram of a sequence is the magnitude of the time-dependent Fourier transform versus time. In a spectrogram a reddish hue indicates high signal strength, where blue indicates low signal strength. Spectrogram of the clean, noisy (degraded by 10dB noise) and enhanced speech signals are presented in Fig. 4.11 for male speakers and in Fig. 4.12 for female speakers. It is clearly visible for noisy signal (Fig. 4.11(b) and 4.12(b)) that, the reddish yellow marks are evenly spread all over the spectrum indicating wide-band WGN. For enhanced these marks are mostly removed. While doing so most enhancement methods crops some part of the actual signal. Fig. 4.11(f), 4.11(g), 4.12(f), 4.12(g) shows that proposed methods removes the background noise significantly while retaining signal details more than other methods.

Table 4.5: Average Itakura-Saito Distortion Measure for various noise levels, obtained using MPE, PARA, MMSE-LSAE and proposed two methods

Speaker	Input SNR [dB]	Degraded IS	Improved IS Measure for Methods				
			MPE	PARA	MMSE-LSAE	Prop. Method	
						Pert1	Pert2
Male	-5	5.78	4.21	4.33	4.19	5.50	5.47
	0	5.03	3.38	3.48	3.70	5.48	5.47
	5	4.22	2.69	2.77	3.28	4.39	4.64
	10	3.45	2.12	2.17	2.76	2.37	2.73
	15	2.76	1.64	1.69	2.16	1.78	2.03
	20	2.17	1.25	1.28	1.62	1.30	1.48
	25	1.67	0.90	0.94	1.29	0.99	1.17
	30	1.25	0.62	0.65	1.28	0.74	0.89
Female	-5	4.69	3.06	3.16	3.33	5.15	5.11
	0	4.00	2.34	2.41	3.02	5.37	5.31
	5	3.24	1.75	1.78	2.71	4.53	4.86
	10	2.49	1.31	1.30	2.43	2.31	2.64
	15	1.82	0.97	0.94	2.13	1.59	1.88
	20	1.28	0.69	0.65	1.71	0.99	1.32
	25	0.85	0.46	0.44	1.22	0.63	1.01
	30	0.54	0.28	0.27	0.81	0.40	0.78

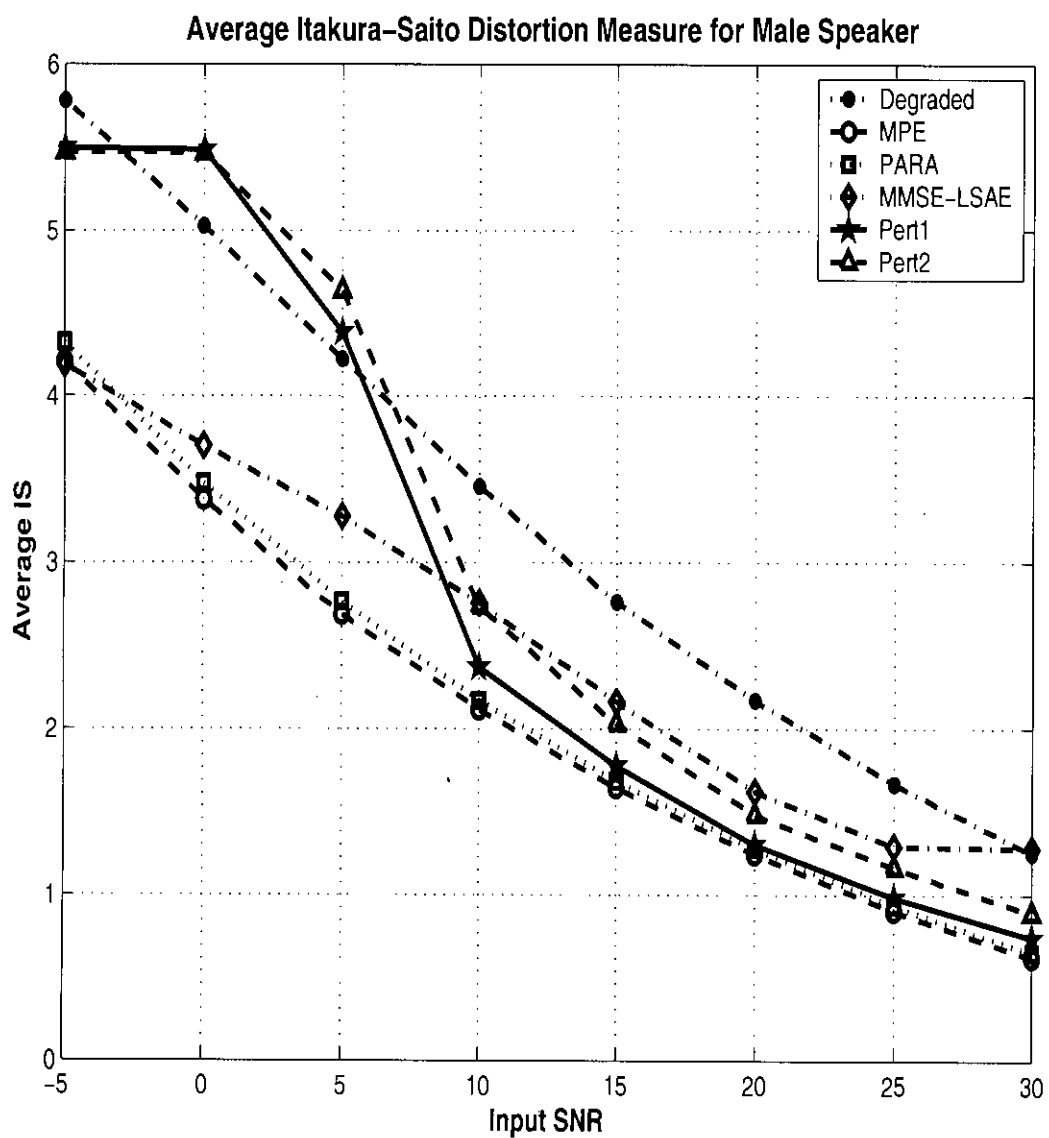


Figure 4.9: Itakura-Saito distortion measure for male utterance: result is averaged using results of 20 utterances

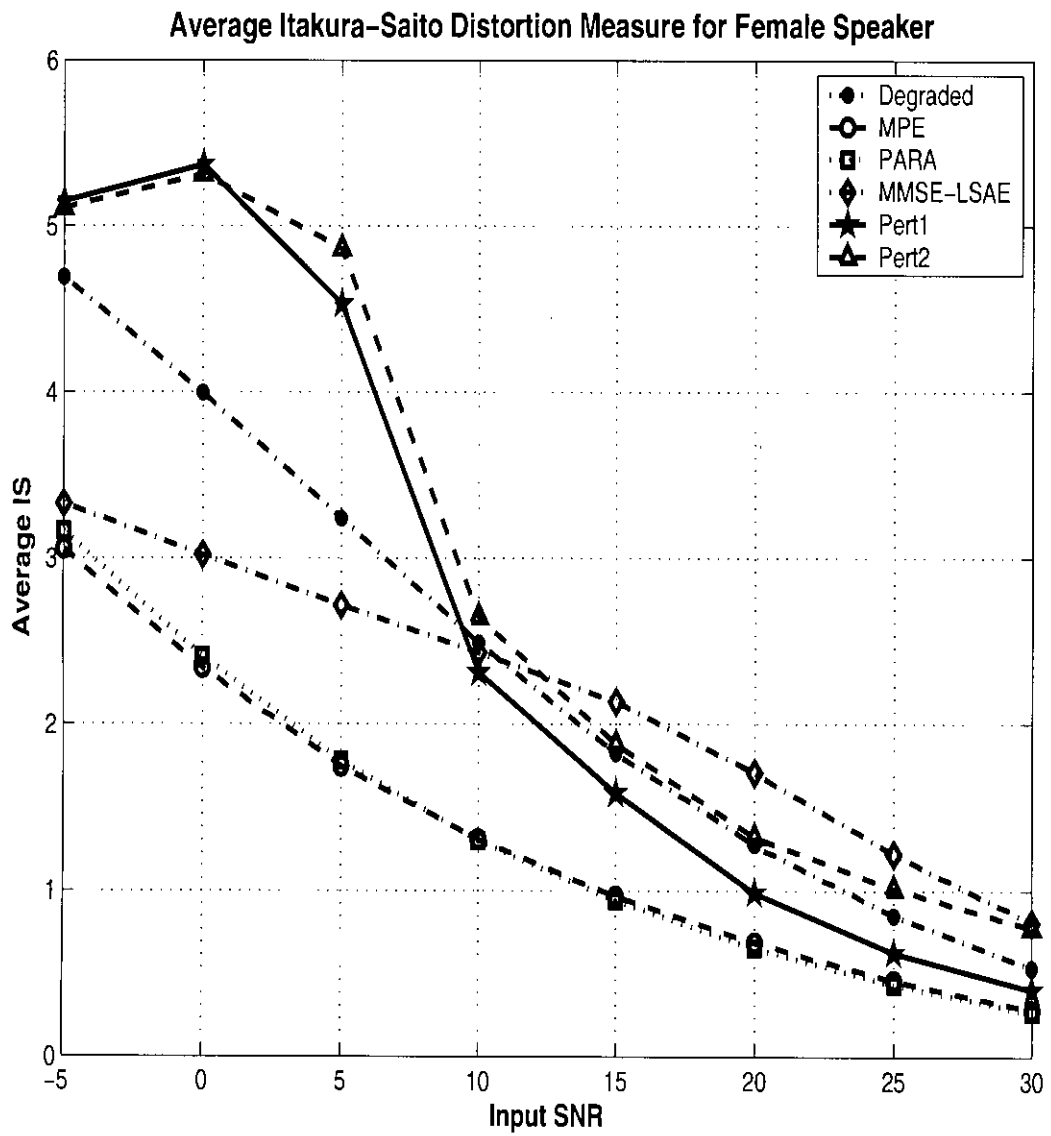


Figure 4.10: Itakura-Saito distortion measure for female utterance: result is averaged using results of 20 utterances

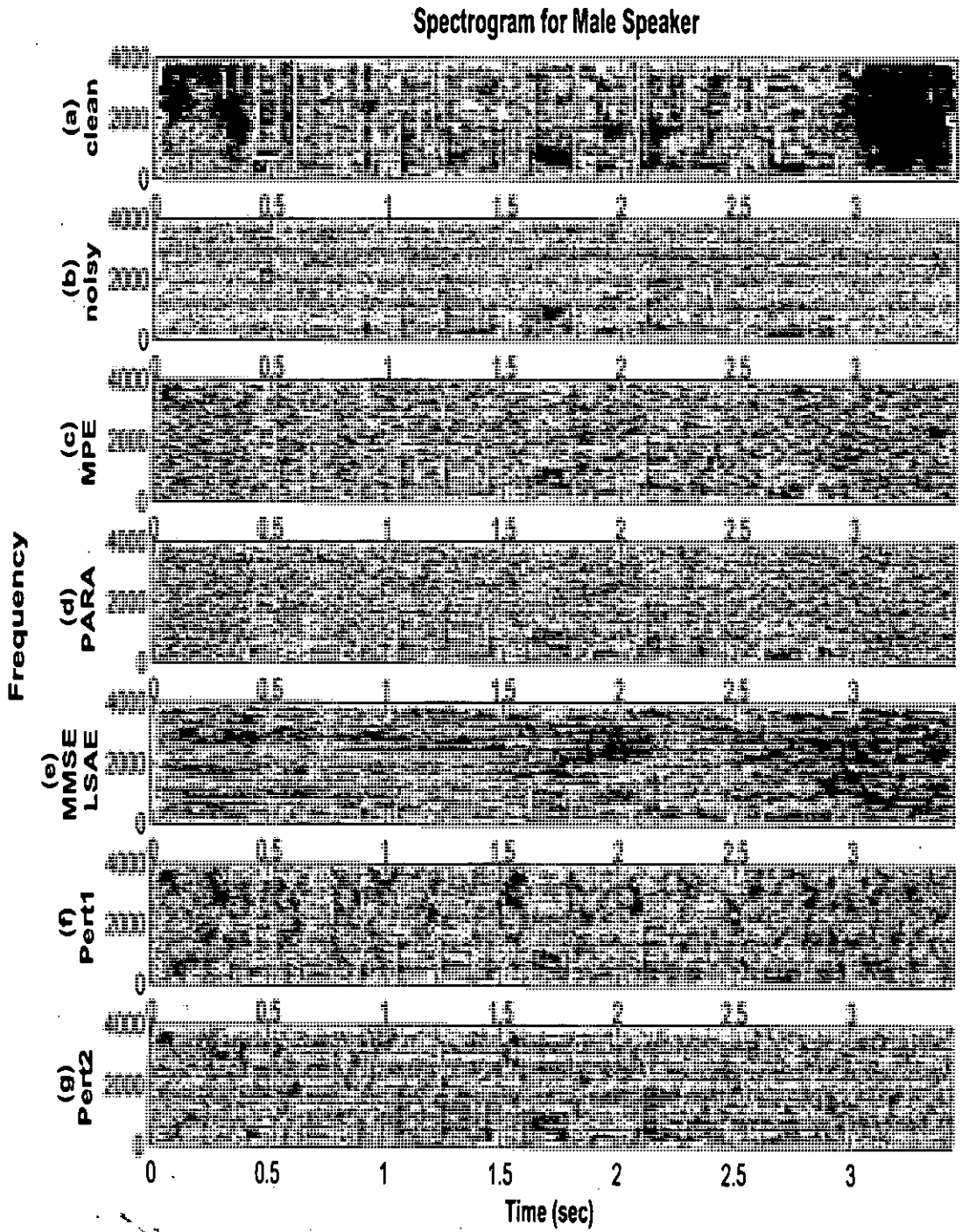


Figure 4.11: Spectrogram for a male utterance "Would you please confirm the government policy regarding waste removal": (a) clean; (b) noisy (corrupted by 10dB WGN); enhanced using (c) MPE; (d) PARA; (e) MMSE-LSAE; (f) Proposed Pert1 and (g) Proposed Pert2

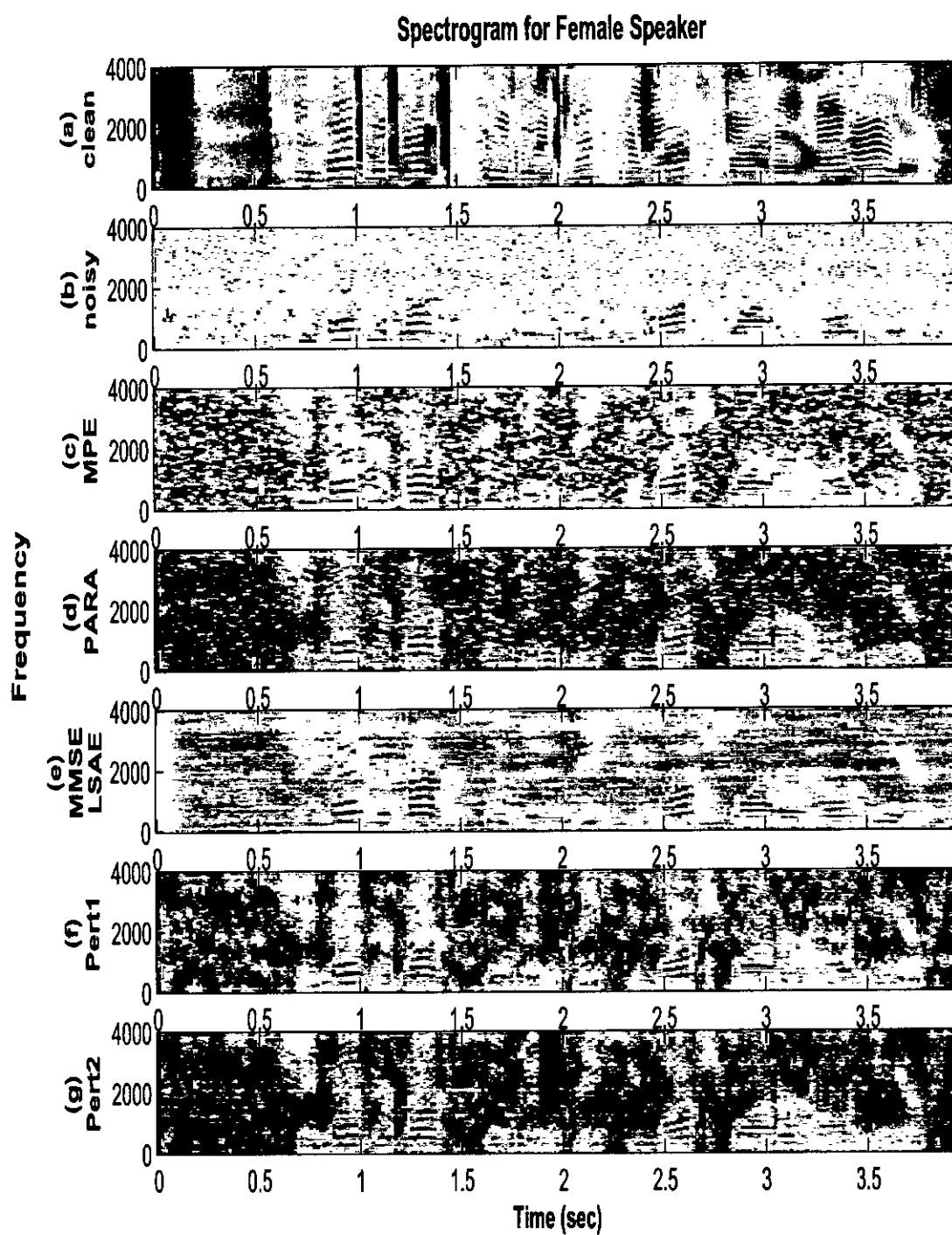


Figure 4.12: Spectrogram for a female utterance “She had your dark suit in greasy wash water all year”: (a) clean; (b) noisy (corrupted by 10dB WGN); enhanced using (c) MPE; (d) PARA; (e) MMSE-LSAE; (f) Proposed Pert1 and (g) Proposed Pert2

4.6 Subjective Test

Considering noise reduction, it is a general notion to think of improving a signal-to-noise ratio (SNR). This may not be the most appropriate performance criterion for speech enhancement. All listeners have an intuitive understanding of speech quality, intelligibility and listener fatigue. However, these areas are not easy to quantify. For subjective quality testing informal listening has been used involving 5 listeners. They voted proposed methods to be better than the MPE for most of the utterances especially for male speakers. For some female utterances listeners could not differentiate between MPE and proposed methods.

4.7 Conclusion

In this chapter various results have been reported. The results for objective and subjective tests of the proposed method are shown along with the results of the three methods reported earlier [35, 36]. Simulation results indicates that the perturbation technique may be a promising one as a new technique for speech enhancement.

Chapter 5

Conclusion

5.1 Summary

A new technique for speech enhancement in DCT domain has been proposed. The major focus of this research was to exploit perturbation as a new technique for speech enhancement. A small perturbation is applied in DCT domain. The amount of perturbation is chosen to be proportional to the average magnitude of noise with respect to estimated clean and noise signal content in DCT domain. For this average trend-line both for noise and signal is first determined. The trend-line for noise is chosen to be constant for a particular time frame and that is the average minima in the DCT domain. And signal trend-line is then obtained by subtracting this noise trend-line from noisy trend-line, forcing negative values to zero. The result of application of perturbation is then compared in time domain with respect to actual noisy signal, to give an estimation of noise distribution. The noise thus estimated is multiplied by a constant (α) such to minimize the error between actual and estimated signal, to subtract it from noisy signal in time domain to have enhanced speech signal.

The result shown in chapter 4 indicates the proposed technique performs better for male speaker, rather than female speakers. Overall SNR, Segmental SNR for proposed 1st method (Pert1) is better than MPE, PARA and MMSE-LSAE for most SNRs. But Pert2 did not perform good for high SNRs. Also Itakura-Saito distortion measure for proposed algorithms (both Pert1 and Pert2) is not better than MPE. But Pert2 will work better for changing noise condition as noise is estimated from the noisy signal of that particular frame. Another thing is that proposed methods is less expensive in terms of processing time, which suggest that the proposed method can be used for real time speech processing

applications.

5.2 Limitations and Suggestion for Future Work

For high SNR the magnitude of noise is negligible with respect to that of signal, that results unintentional change in speech while applying perturbation on noise. Consequently improvement is not that good for high SNRs. Also the proposed method does not produce better results for female voice. So a better perturbation criteria needs to be investigated for high input SNRs as well as for female voices.

The perturbation technique ensures better intelligibility, which is evident from objective tests, also the quality of enhanced speech as estimated from subjective test give better enhancement quality. But still there are some musical noise present in enhanced speech. Masking properties of human auditory system [17] may be combined with the proposed methods to improve the performance.

The value β (Eq. 3.7) is chosen empirically, which may be investigated further to obtain an optimized value to give better performance in terms of intelligibility and quality. Also research is needed to include various real and colored noise.

Bibliography

- [1] 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.
- [2] E. Visser, T. W. Lee and M. Otsuka, "Speech Enhancement in Car Environments," in *Proc. 3^d International Conference on Independent Component Analysis and Source Separation*, pp. 272-277, San Diego, December 2001
- [3] N. D. Degan and 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.
- [4] 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.
- [5] J. H. L. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *ICSLP-98: Inter. Conf. on Spoken Language Processing*, vol. 7, pp. 2819-2822, Sydney, Australia, Dec. 1998
- [6] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 113-120, 1979.
- [7] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586-1604, 1979.
- [8] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. of IEEE ICASSP*, pp. 208-211, Washington DC, 1979.

- [9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Speech Audio Processing*, vol. ASSP-32, pp. 1109-1121, 1984.
- [10] R. Hoeldrich and M. Lorber, "Broadband noise reduction based on spectral subtraction," *Proc. ICSPAT*, pp. 265-269, 1997.
- [11] P. Scalart and J. Vieira Filho, "Speech enhancement based on a priori signal to noise estimation," *Proc. ICASSP*, pp. 629-632, 1996.
- [12] R. McAulay and M. Malpass, "Speech enhancement using a soft decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, pp. 137-145, 1980.
- [13] B. L. Sim, Y. C. Tong, J. S. Chang and C. T. Tan, "A Parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 328-337, 1998.
- [14] Z. Nemer, R. Goubran and S. Mahmoud, "Speech enhancement using fourthorder cumulants and optimum filters in the subband domain," *Speech Communication*, vol. 36, pp. 219-246, 2002.
- [15] C. Avendano, H. Hermansky, M. Vis and A. Bayya, "Adaptive speech enhancement using frequency specific SNR estimates," *Proceedings of III IEEE Workshop on Interactive voice Technology for Telecommunications Applications*, Basking Ridge, New Jersey, pp. 65-68, 1996.
- [16] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, issue: 4, pp. 679-681, 1982.
- [17] N. Virag, "Single channel speech enhancement system based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, issue: 2, pp. 126-137, 1999.
- [18] 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.
- [19] 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.

- [20] 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, 1998, pp. 165-181.
- [21] D. L. Donoho, "Denoising by soft thresholding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 613-627, May 1995.
- [22] 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.
- [23] E. Ambikairajah, G. Tattersall and A. Davis, "Wavelet transform-based speech enhancement," in *Proceedings of ICSLP*, 1998.
- [24] Y. Ephraim, "Statistical model based speech enhancement systems," *Proc. of IEEE*, vol. 80, issue. 10, pp. 1526-1555, 1992,
- [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] M. J. Borran and R. D. Nowak, "Wavelet-based denoising using hidden Markov models," in the Proceeding of *ICASSP 2001*.
- [28] B. Yegnanarayana, C. Avendano, H. Hermansky and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, pp. 25-42, 1999.
- [29] J. S. Lim and A. V. Oppenheim, "All pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP 26, pp. 197-210, 1978.
- [30] 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.

- [31] 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.
- [32] I. Y. Soon and S. N. Koh, "Low distortion speech enhancement," *IEE Proc. Vis. Image Signal Process.*, vol. 147, pp. 247-253, 2000.
- [33] I. Y. Soon, S. N. Koh and C. K. Yeo, "Improved noise suppression filter using self adaptive estimator of probability of speech absence," *Signal Processing*, vol. 75, no. 2, pp. 151-159, Jun 1999.
- [34] S. Salahuddin, S. Z. A. Islam, M. K. Hasan and M. R. Khan, "Speech enhancement by envelope restoration and soft thresholding in DCT domain," *Proceedings of ICECE 2002*, pp. 156-159, Dhaka, Bangladesh, 26-28 Dec 2002.
- [35] M. K. Hasan, S. Salahuddin and M. R. Khan, "A modified *A priori* SNR for speech enhancement using spectral subtraction rules," *IEEE signal processing letters*, vol. 11, no. 4, pp. 450-453, April 2004.
- [36] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.
- [37] 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.
- [38] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Transactions on Speech and Audio Processing*, vol. 7, issue: 1, pp. 1 -17, 1999.
- [39] A. Kumar and S. K. Mullick, "Nonlinear Dynamical Analysis of Speech," *Journal of the Acoustical Society of America*, vol. 100, pp. 615-629, 1996.
- [40] H. D. I. Abarbanel, *Analysis of observed chaotic data*, New York: Springer, 1996.
- [41] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of Statistical Physics*, vol. 65, pp. 579-616, 1991.

- [42] H. Kantz and T. Schreiber, *Nonlinear time series analysis*, Cambridge: Cambridge University Press, 1997.
- [43] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, pp. 504-512, Vol. 9, No. 5, July 2001.
- [44] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. 7th EUSIPCO'94*, Edinburgh, U.K., Sept. 13-16, 1994, pp. 1182-1185.
- [45] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, pp. 12-15, VOL. 9, NO. 1, January 2002
- [46] S. Rangachari, P.C. Loizou and Yi Hu, "Noise estimation algorithm with rapid adaptation for highly nonstationary environments," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. (ICASSP '04)*, pp. 305-308, Vol. 1, 17-21 May 2004.
- [47] Sir M. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Great Britain: Charles Griffin & Company Ltd., vol. 1, 4th ed., 1977.
- [48] S.R. Quackenbush, T.P. Barnwell and M.A. Clements, *Objective Measures of Speech Quality*, NJ: Prentice-Hall, 1988.