Speech Enhancement by Modeling of Stationary Time-Frequency Regions

By

Rubén E. Galarza

B.S., University of Puerto Rico – Mayagüez (1996)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering

at the

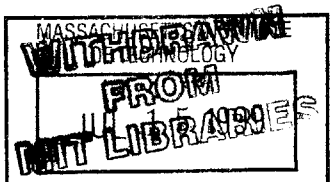
Massachusetts Institute of Technology

June 1999

Author_____
Department of Electrical Engineering and Computer Science
May 7, 1999

Certified by_____
Jae S. Lim
Professor of Electrical Engineering
Thesis Supervisor

Accepted by_____
Arthur C. Smith
Chairman, Committee on Graduate Students
Department of Electrical Engineering and Computer Science



Speech Enhancement by Modeling of Stationary Time-Frequency Regions

By
Rubén E. Galarza

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Master of Science in Electrical Engineering

Abstract

The main objective of speech enhancement is to improve the overall quality and intelligibility of degraded speech. Speech enhancement has been studied for many years, and numerous enhancement algorithms have been developed. However, these traditional techniques tend to improve the signal-to-noise ratio (SNR) of the signal at the expense of losing some intelligibility in the speech.

Recently, a new enhancement technique was developed which achieves a better compromise between improved SNR and intelligibility. The algorithm appears to be successful by identifying and modeling stationary regions in the time-frequency plane. This time-frequency segmentation permits the algorithm to bypass certain problems encountered in traditional enhancement algorithms.

This thesis deals with the overall improvement of this enhancement system, in terms of computational efficiency and the quality and/or intelligibility of its output. The procedure was simplified by generalizing some of its modeling assumptions. The enhanced speech quality and intelligibility was improved by relaxing the algorithm's modeling constraints at high frequency regions. Informal tests showed that these modifications provided some perceptual improvements compared to previous incarnations of the algorithm.

Thesis Supervisor: Jae S. Lim
Title: Professor of Electrical Engineering

Dedication

To

My Parents

Acknowledgements

I want to take this opportunity to thank all the people that helped me both professionally and personally during my stay at MIT. First, I would like to thank my advisor Dr. Jae S. Lim, for his unconditional support, encouragement, and most of all, for his un-dying patience. I would like to thank Dr. Chang Dong Yoo for all his support and advice. Without his help, it is safe to say I wouldn't have a thesis today. Likewise, I want to thank the people from my research group: David Baylon, Ray Hinds, Theresa Huang, Eric Reed, Wade Wan and Cindy LeBlanc. All of them were instrumental in my adjustment to graduate life at MIT, and were always ready and willing to help when I needed them. Thank you all.

On a more personal note, I would like to thank my housemates Luis Tomas Perez-Prado, and Iris Ortiz. Although I've known them for less than a year, they have become my family in Boston. Thanks for your support and your love. I want to thank my dear friend Orlando Navarro. He was far away physically, but very close to me in spirit. Thanks for your advice "Obi-Wan". I also want to send my deepest love to my previous housemates Israel Almodovar, and Victor Valentín. I was going through some rough times when we moved in together, but they reminded me how to enjoy life again. We shared some of the best moments of my life together. I will always be grateful for their friendship. Ustedes son los hermanos mayores que nunca tuve.

Finally, I would like to thank my family: Mami, Papi and my little sister Rocio. I can't thank God enough for putting me in this wonderful family. In the worst of times, when it felt like everybody had turned away from me, they were there to help me get through it. Without their support, advice and understanding I never would've finished this. Their mere existence defines love. Los amo siempre. Gracias.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Degradation of speech due to additive noise occurs in many types of situations. Disturbances of this type may vary from low-level office noise in a normal phone conversation to high volume engine noise in a helicopter or airplane. In general, additive noise reduces intelligibility and introduces listener fatigue. Speech degraded by noise also affects the performance of speech recognition and speech coding systems, which may have been developed assuming a noise-free speech input. For these reasons, enhancing speech is desirable.

Speech enhancement has been an active area of research for more than 20 years. Its main objective is to improve perceptual aspects of degraded speech signals, such as quality and intelligibility. Various enhancement methods have been developed over the years, each with its benefits and disadvantages. A common problem of traditional methods is the introduction of substantial artifacts in the processed speech.

The ideal enhancement procedure would maximally reduce noise while minimizing the distortion and artifacts added to the signal. The usual approach of enhancement algorithms is to seek a compromise between these opposing goals over the entire spectrum of fixed length segments in the signal. However, a new enhancement technique introduces the idea of exploiting local time-frequency characteristics of speech to achieve a better trade-off. By segmenting speech into stationary regions in time-frequency, the stationarity constraints posed by traditional enhancement algorithms are met and better quality speech is produced.

To achieve an even better compromise between quality and noise reduction, this new enhancement system was modified to identify and model general characteristics of speech. In this manner, better models of each time-frequency region were obtained and more of the original speech came through without any added artifacts. This thesis presents a detailed look of this speech enhancement procedure as well as the various modifications made to simplify it and obtain better quality speech.

1.1 Problem

The problem of enhancing speech degraded by noise is seen in a large number of scenarios. This makes it necessary for researchers to establish a number of assumptions before the enhancement system is developed. For example, they must define the type of noise interference, the way the noise interacts with the speech signal, and the number of channels available for enhancement. Different types of noise include competing speakers, background sounds from an office environment, traffic, wind or random channel noise. The noise might affect the original signal in an additive, convolutional or multiplicative manner and it may be dependent or independent of the original speech. Also, more than one channel of speech information could be used in the enhancement process, since auxiliary microphones can be employed to monitor the noise source while others monitor the degraded speech.

This thesis studies the specific scenario of speech degraded by additive noise with only a single channel of information available. The degraded signal is expressed as

$$y[n] = s[n] + z[n] \tag{1.1}$$

where s[n] is a speech sequence, and z[n] is a noise sequence which is independent of s[n]. Another assumption is that the sequence z[n] represents stationary noise. That is, the noise is assumed to maintain the same characteristics throughout the duration of the signal (as opposed to non-stationary noise whose characteristics change throughout the

duration of the signal.) Also, the noise power spectrum is assumed known and can be estimated using segments of no speech activity within the signal. For simplicity, it will be assumed that the noise power spectrum is white, but the results extend to colored background noise.

1.2 Objective

Speech enhancement in general has the objective of improving the overall quality and intelligibility of degraded speech. Speech quality is a subjective measure reflected in the way the signal is perceived by the listeners [2]. It is related to the pleasantness of the signal sound or to the amount of effort incurred by the listener to understand the message. On the other hand, intelligibility is the amount of information that can actually be extracted from the speech signal by the listener. Although both measurements are related, they are not exactly the same, since a lot of information could be extracted from a speech signal, even if a lot of effort is needed to extract it.

Noise reduction is thought of as an improvement in the signal-to-noise ratio (SNR) of a given signal. An improvement in SNR tends to increase the quality of the degraded speech, but it does not guarantee an increase in intelligibility. Enhancement systems improve the SNR of the speech signal, but their output tends to have reduced intelligibility because some aspects of the original speech are lost in the process. Furthermore, these systems usually add artifacts that reduce the quality of the signal, regardless of any SNR improvement. The systems presented in the following chapters are superior to more traditional enhancement procedures in that they remove less original speech and produce no artifacts in the speech signal.

In general, intelligibility and quality are hard to quantify. For this reason, the evaluation of the presented systems will be primarily based on informal listening. Other more objective measurements like segmental SNR will also be used to evaluate the system.

1.3 Outline of Thesis

Chapter 2 presents a brief overview of traditional speech enhancement systems. By understanding the strengths and weaknesses of each method, it is easier to see how the new systems achieve better results.

Chapter 3 describes a novel speech enhancement system developed by Dr. Chang Dong Yoo in 1996. This system is different from traditional techniques in that it uses filtering and adaptive windowing to divide the signal into stationary time-frequency regions. The chapter explains the algorithm used for segmenting the signal into frequency bands. It also describes how a varying length window is produced following the spectral characteristics of the speech signal. Finally, it demonstrates how each region is modeled and enhanced by selective linear prediction and modified Wiener filtering.

Chapter 4 discusses the modifications made to the enhancement system presented in Chapter 3. These changes were made primarily in the enhancement stages of the algorithm. They were implemented to improve the computational efficiency of the algorithm and the intelligibility of the enhanced speech.

Chapter 5 describes the performance of the modified speech enhancement system. Segmental SNR and Itakura-Saito measurements are presented as well as results obtained from subjective informal listening.

Finally, Chapter 6 presents a summary of the basic ideas presented. In addition, it includes suggestions for future research.

Chapter 2

Traditional Enhancement Techniques

2.1 Introduction

Various speech enhancement algorithms have been developed over the years. In general, they are capable of producing acceptable speech for specific applications. However, they also tend to introduce artifacts in the speech signal. Understanding the strengths and weaknesses of some of these procedures can provide valuable insight in the study of new speech enhancement algorithms. Therefore, a brief overview of these traditional techniques is in order.

There are two major types of single channel enhancement systems: those concentrated on the short-term spectral domain, and those based on different models of speech. The short-term spectral domain algorithms find an estimate of the noise bias in the degraded speech and subtract it to produce an enhanced version. The model based algorithms focus on estimating clean speech model parameters from the degraded signal, so that the problem of enhancing speech becomes one of parameter estimation. A brief description of these two types of systems is now presented.

2.2 Short-Time Spectral Subtraction

Short-time spectral subtraction is based on subtracting a noise spectral density estimate from the degraded signal to obtain an enhanced signal. The subtraction is performed on a frame by frame basis, where each frame usually consists of windowed speech 20 to 40ms in duration. The analysis window has a fixed length throughout the process and the subtraction is implemented in the power spectrum, Discrete-Time Fourier

Transform (DTFT) or auto-correlation domain. The noise spectral density is estimated by using areas of non-speech activity in the signal.

The windowed speech $y_w[n]$ is defined by

$$y_w[n] = s_w[n] + z_w[n] \tag{2.1}$$

where the subscript *w* indicates that the signal is obtained by applying a window function w[n] to the degraded speech y[n] (i.e. $y_w[n] = w[n] \cdot y[n]$). The window is shifted in time as other segments of the signal are analyzed.

From Equation (2.1), an estimate of the clean speech's short-time spectral magnitude is found by

$$\left|\hat{S}_w(\omega)\right|^2 = \left|Y_w(\omega)\right|^2 - E\left[\left|Z_w(\omega)\right|^2\right] \tag{2.2}$$

where $Y_w(\omega)$ and $Z_w(\omega)$ are the Fourier transforms of $y_w[n]$ and $z_w[n]$ respectively. Since $|Z_w(\omega)|^2$ is not directly available, it is estimated by $E[|Z_w(\omega)|^2]$, where $E[\bullet]$ represents the ensemble average.

The estimate $|\hat{S}_w(\omega)|$ can be generalized by

$$\left|\hat{S}_w(\omega)\right|^a = \left|Y_w(\omega)\right|^a - c \cdot E\left[\left|Z_w(\omega)\right|^a\right] \tag{2.3}$$

where constants a and c represent extra degrees of freedom used to enhance the algorithm performance. The windowed speech estimate $\hat{S}_w[n]$ is obtained by combining the magnitude $|\hat{S}_w(\omega)|$ with the phase of the noisy signal $\angle Y_w(\omega)$.

The most popular example of these techniques is called spectral subtraction. Figure 2.1 shows the implementation of this enhancement algorithm.

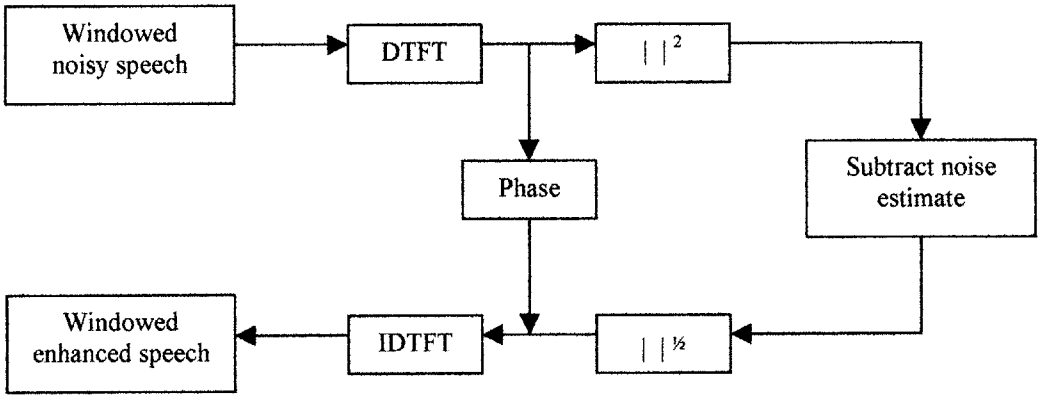


Figure 2.1: Algorithm for spectral subtraction.

The figure shows how the subtraction of the noise bias takes place in the power spectrum domain. In general, these subtraction techniques follow the same scheme. Their main difference lies in the domain in which the subtraction takes place.

Short-time spectral subtraction techniques are useful in many applications due to their general lack of complexity. However, they are not the most effective enhancement procedures, since they tend to de-emphasize unvoiced speech and high frequency formants. Furthermore, they introduce "musical tones" in the enhanced signal [1, 2], a problem that leads to listener fatigue.

2.3 Model Based Enhancement Systems

The second class of speech enhancement techniques includes systems based on various speech models. These methods focus on estimating clean speech model parameters from the degraded signal. They are generally considered more effective than subtraction techniques, although each technique is limited by the underlying assumptions of its model. Some examples presented here are the statistical model system, the dual excitation (DE) model system, and the maximum *a posteriori* (MAP) system.

2.3.1 MAP Based Method

Figure 2.2 shows a traditional discrete model for short-time speech production. This model represents speech as the output of a filter excited by a sequence with two possible states. The excitation sequence $u_w[n]$ is modeled either by a periodic pulse train for voiced speech or by random noise for unvoiced speech. The filter H(z) represents the vocal track of the speaker and it is modeled with an all-pole transfer function given by

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k \cdot z^{-k}} \tag{2.4}$$

where G, p and a_i are the gain, order and linear prediction coefficients of the all-pole model.

From Equation 2.4 and Figure 2.2 a difference equation for the all-pole model can be derived:

$$s_w[n] = \sum_{k=1}^{p} a_k \cdot s_w[n-k] + G \cdot u_w[n]. \tag{2.5}$$

Equation 2.5 shows how each sample of the speech signal $s_w[n]$ is obtained from the excitation input $u_w[n]$ and previous speech samples weighted by the corresponding linear prediction coefficients.

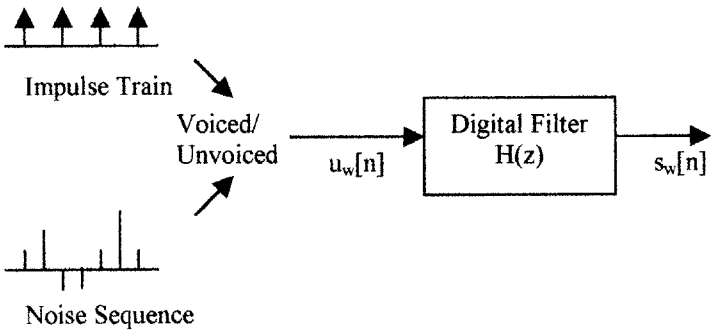


Figure 2.2: Traditional speech production model.

The linear prediction coefficients a_i and the gain G are estimated from the noise free signal by solving a set of linear equations. This process is known as the autocorrelation or covariance method, depending on the set of initial conditions assumed [3]. On the other hand, solving for these parameters in the presence of background noise produces a set of non-linear equations, demanding substantial computational resources. An alternate sub-optimal solution for the noisy speech case is presented in [4]. This solution simplifies to the iterative two step algorithm presented in Figure 2.3. First, the linear prediction coefficients are initialized by using the auto-correlation method on the degraded speech. Then, the coefficients are used in a non-causal Wiener filter to enhance the degraded signal. Finally, the output of the filtering process is used to calculate a new set of linear prediction coefficients. The last two steps are repeated for a few iterations to obtain an enhanced signal. The non-causal Wiener filter has the following frequency response [5]

$$H_w(\omega) = \frac{P_s(\omega)}{P_s(\omega) + c \cdot P_z(\omega)} \tag{2.6}$$

where $P_s(\omega)$ is given by

$$P_s(\omega) = \frac{G^2}{\left|1 - \sum_{k=1}^{p} a_k \cdot e^{(-j\omega k)}\right|^2}. \qquad (2.7)$$

Also, $P_z(\omega)$ is an estimate of the noise power spectrum and c is a constant that provides the Wiener filter with an extra degree of freedom .

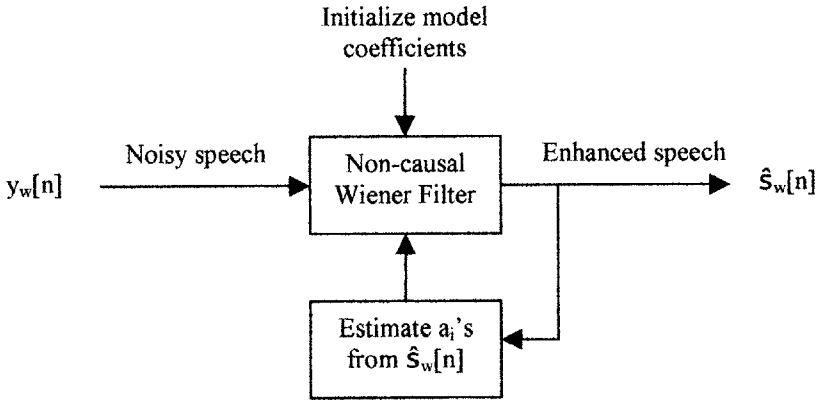


Figure 2.3: MAP algorithm.

This algorithm has been shown to increase the joint likelihood of the speech sequence and the all-pole parameters after each iteration. Its main drawback is that all-pole models are biased towards frequency components of high energy. In addition, the speech within the analysis interval is assumed stationary during the estimation of the model parameters. For an analysis window of fixed length, this assumption is false since stationarity varies for different speech sounds and different speakers.

2.3.2 Statistical Model-Based Speech Enhancement

In the statistical model-based speech enhancement method, the windowed speech segment $s_w[n]$ is modeled as sound generated by one of a finite number of sub-sources. Each sub-source represents a particular class of statistically similar speech sounds with a particular power spectrum, a parametric probability distribution function $p_{\lambda_S}(s)$ and an autoregressive process of a given order.

Speech segments are classified into different categories, each with its corresponding estimator. Drucker [6] classifies segments into five categories: fricatives, stops, vowels, glides, and nasals, as shown in Figure 2.4. Depending on how the degraded speech segment is classified, different enhancing filters are used to remove the noise. Another algorithm developed by McAulay and Malpass [7] classifies segments as part of silent or non-silent states. A weighted sum of estimators for both states produces the enhanced segments. The weighting coefficients are obtained from *a posteriori* probabilities of the states given the noisy speech signal.

All sub-sources represent acoustic signals generated from a fixed configuration of the vocal track. Each configuration is modeled as an all-pole filter. The transition from one sub-source to another is modeled in a Markovian manner. Since these transitions are hidden from the listener, the model is referred to as a Hidden Markov Model (HMM). In addition, each sub-source has a unique spectral prototype for speech and noise, so Wiener filters can be designed and used according to the error criterion and the probability distribution of each prototype.

The main difference between this method and the others mentioned so far is that the speech model parameters are estimated from training data of clean speech, instead of estimating them directly from the degraded speech signal. However, this also proves to be one of the method's biggest drawbacks, since it requires extensive "training" to estimate the statistical parameters involved in the HMM. In addition, the recording

conditions of the test and training data must be similar. The algorithm also suffers from the same limitations of the MAP method, since it employs an all-pole model of speech.

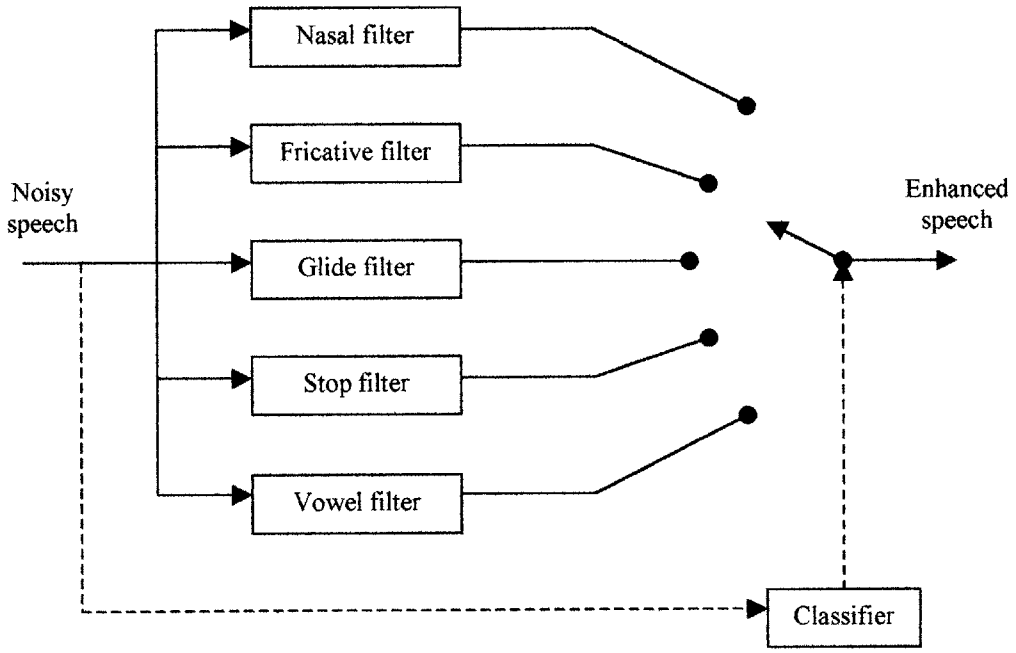


Figure 2.4: "Five sub-sources" approach for statistical-model-based speech enhancement

2.3.3 Enhancement using the Dual-Excitation Model

The dual excitation (DE) model overcomes some of the limitations of other speech models by decomposing the windowed speech signal $s_w[n]$ into co-existing voiced and unvoiced components. The voiced component is denoted by $v_w[n]$ and the unvoiced component by $u_w[n]$. In the Fourier domain, the speech signal is expressed as

$$S_w[\omega] = V_w[\omega] + U_w[\omega] \tag{2.8}$$

where $S_w(\omega)$, $V_w(\omega)$, and $U_w(\omega)$ are the DTFT's of $s_w[n]$, $v_w[n]$ and $u_w[n]$ respectively.

Assuming the voiced component is periodic over the duration of the analysis window, its pitch period can be used to model it as a harmonic series. Thus, $V_w(\omega)$ is a sum of various harmonics of the fundamental frequency ω_0. Following this assumption, the mathematical expressions for $v_w[n]$ and $V_w(\omega)$ are

$$v_w[n] = \sum_{n=-M}^{M} A_m w[n] e^{-jnmw_0} \tag{2.9}$$

$$V_w[\omega] = \sum_{m=-M}^{M} A_m W(\omega - mw_0) \tag{2.10}$$

where $W(\omega)$ is the Fourier Transform of the window function w[n] and A_m represents the amplitude of the m^{th} harmonic. $W(\omega)$ is essentially a narrow band low-pass filter.

Since the DE model parameters are not known, they must be estimated from the speech spectrum. The estimates of the fundamental frequency and the harmonic amplitudes are obtained with an algorithm developed by Griffin [8]. This algorithm minimizes the mean squared error between the original speech spectrum $S_w(\omega)$ and the voiced speech spectrum $V_w(\omega)$ and ensures that the voiced component will contain all the harmonic structure of the original speech.

The unvoiced spectrum $U_w(\omega)$ is estimated from the difference spectrum $D_w(\omega)$ defined as

$$D_w[\omega] = S_w[\omega] - V_w[\omega]. \tag{2.11}$$

In general, different types of smoothing are used on $D_w(\omega)$'s spectral magnitude to obtain the magnitude of the unvoiced component $|U_w(\omega)|$. This is done under the assumption

that the fine structure of $|U_w(\omega)|$ doesn't need to be completely preserved due to the inherent characteristics of unvoiced sounds. The phase of the unvoiced component $\angle U_w(\omega)$ is often set equal to the phase of the difference spectrum or to the phase of the reference noise signal.

The enhancement of the speech signal is performed on the voiced and unvoiced components separately. Enhancement of the voiced component entails only the modification of the harmonic amplitudes since the estimation error of the fundamental frequency ω_0 is assumed negligible. To modify the harmonic amplitudes A_m, the basic algorithm is to eliminate the m^{th} amplitude estimate if its value is less than the value of effective noise at the corresponding frequency. The enhanced harmonic amplitudes $\hat{A}m$ are given by

$$\hat{A}_m = \begin{cases} 0 & if \ |A_m| < 3\left[\dfrac{P_{zz}(m\omega_0)}{N_{eff}}\right]^{\frac{1}{2}} \\ A_m & otherwise \end{cases} \tag{2.12}$$

where $P_{zz}(\omega)$ represents the noise power density. The parameter N_{eff} accounts for windowing effects and is defined as

$$N_{eff} = \frac{\left[\sum_{n=-\infty}^{\infty} w^2[n]\right]^2}{\sum_{n=-\infty}^{\infty} w^4[n]} \tag{2.13}$$

The enhancement of the unvoiced component $U_w(\omega)$ is a two step process. First, the harmonic bands where the voiced energy is substantially greater than the unvoiced

energy are identified in the difference spectrum. In these bands, the voiced energy masks the unvoiced energy. Therefore, the unvoiced energy can be eliminated without altering the perceived speech. The enhanced difference spectrum $\check{D}_W(\omega)$ is defined as

$$\check{D}_W(\omega) = \begin{cases} 0 & if\ E_{v_m} > 3E_{uv_m} \\ D_W(\omega) & otherwise \end{cases} \tag{2.14}$$

where E_{v_m} and E_{uv_m} are the energies of the voiced and unvoiced components at the m^{th} harmonic. In the second step, a modified Wiener filter is applied on $\check{D}_W(\omega)$ to remove residual background noise where the spectrum has a low signal to noise ratio. The Wiener filter is defined as

$$H_{w_{ss}}(\omega) = \begin{cases} \beta & if\ \dfrac{\alpha \cdot E\left[\left|Z_{w_{ss}}(\omega)\right|^2\right]}{E\left[\left|\check{D}_{w_{ss}}(\omega)\right|^2\right]} > 1 \\ 1.0 - \dfrac{\alpha \cdot E\left[\left|Z_{w_{ss}}(\omega)\right|^2\right]}{E\left[\left|\check{D}_{w_{ss}}(\omega)\right|^2\right]} & otherwise \end{cases} \tag{2.15}$$

where $E[|Z_{w_{ss}}(\omega)|^2]$ is an estimate of the noise power spectrum and $E[|\check{D}_{w_{ss}}(\omega)|^2]$ is the smoothed unvoiced spectrum. The subscript denotes the application of the window function $w_{ss}[n]$. Some typical values for α and β are 1.6 and 1 respectively.

The DE algorithm for speech enhancement has been shown to produce better quality speech than spectral subtraction algorithms. However, this technique also encounters problems due to its model assumptions. Its main drawback is the assumed perfect periodicity of voiced speech. Since voiced speech sounds are only quasi periodic, this assumption leads to inaccurate voiced/unvoiced decompositions.

2.4 Discussion

In this chapter, a brief overview of the more popular systems for speech enhancement was presented. Each technique was described in terms of its functionality, its benefits, and its main drawbacks. Spectral subtraction algorithms are simple and give acceptable results for some applications, but introduce musical tones in the speech signal and tend to de-emphasize high frequency content. Model-based systems perform better in reducing noise than spectral subtraction systems. However, these systems are often inadequate in representing speech due to their underlying model assumptions.

Since model based systems have been more successful, it might be beneficial to think of ways to improve their performance. Recent work proposes that it might be possible to overcome some of the shortcomings of model based systems by dividing speech into stationary regions in the time-frequency plane and then enhancing each region according to its spectral characteristics. This idea is the basis for a novel speech enhancement system developed in [9] that was shown to produce better results than the traditional enhancement systems in this chapter. A detailed description of this system is presented in the next chapter.

Chapter 3

Novel Speech Enhancement System

3.1 Introduction

Model-based systems are the best enhancement methods developed so far. However, they have inherent disadvantages due to inadequacies in their models. To overcome some of their shortcomings, a new approach for speech enhancement was developed by Dr. Chang Dong Yoo in 1996 [2]. This system divides the time-frequency plane of the speech signal into stationary regions and then models each region according to its local spectral characteristics. This chapter will present a detailed description of this algorithm.

3.2 Stationary Region Segmentation

The first step in the enhancement process is the identification of stationary time-frequency regions. The segmentation of the time-frequency plane is achieved by two main procedures: M-band decomposition of the signal, and application of an adaptive analysis window. M-band decomposition separates speech into multiple frequency bands. This is helpful because any variation in the speech characteristics can be isolated to a given set of frequencies. Adaptive windowing allows maximum averaging of speech segments while minimizing temporal smearing. Together, both processes produce longer stationary analysis intervals that follow the spectral characteristics of the speech signal across the time-frequency plane.

3.2.1 M-Band Frequency Segmentation

An M-band decomposition scheme is presented in Figure 3.1(a). The degraded signal y[n] is divided into M channels $y^{(i)}[n]$ such that

$$\sum_{i=1}^{M} y^{(i)}[n] = y[n]. \tag{3.1}$$

Without loss of generality, $y^{(1)}[n]$ is the dc channel while $y^{(k-1)}[n]$ and $y^{(k)}[n]$ are contiguous channels (where $y^{(k)}[n]$ is higher in frequency).

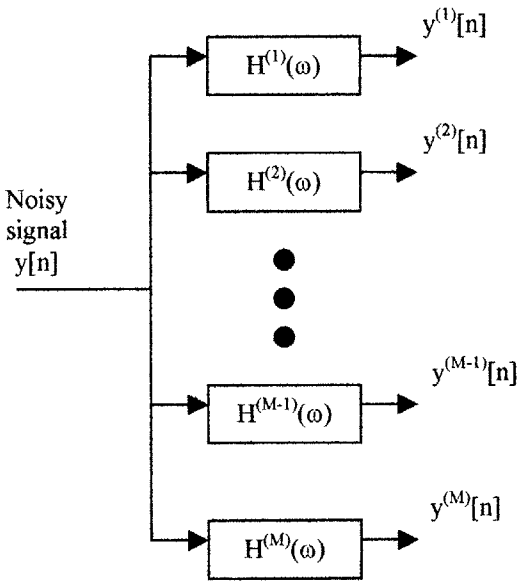
By looking at Figure 3.1(a) it is clear that for Equation 3.1 to hold, the band-pass filters must satisfy the relation

$$\sum_{j=1}^{M} H^{(j)}(\omega) = 1 \tag{3.2}$$

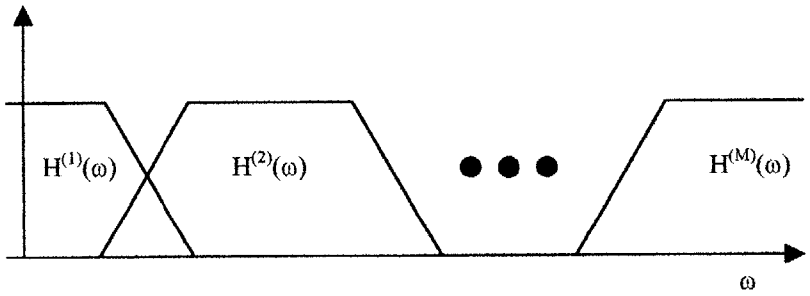
where $H^{(j)}(\omega)$ is the frequency response of the j^{th} band-pass filter. An example is shown in Figure 3.1(b).

There are various ways to achieve the M-band decomposition described above. In the current system, the signal is separated into M frequency channels by a group of M-1 low-pass filters (LPF) connected in series. Figure 3.2 shows how the frequency channels are obtained by subtracting the input and output of each LPF $G^{(i)}(\omega)$. The filters are designed so that any given $G^{(N)}(\omega)$ has a bigger pass-band than the next filter $G^{(N-1)}(\omega)$. The relationship between the low-pass filters $G^{(i)}(\omega)$ and the band-pass filters $H^{(j)}(\omega)$ is given by

$$\prod_{i=j}^{M-1} G^{i}(\omega)[1 - G^{(j-1)}(\omega)] = H^{(j)}(\omega). \tag{3.3}$$



(a)



(b)

Figure 3.1: (a) Filter bank for decomposition of degraded speech into M-bands. (b) The band-pass filters must add up to unity in the spectral domain.

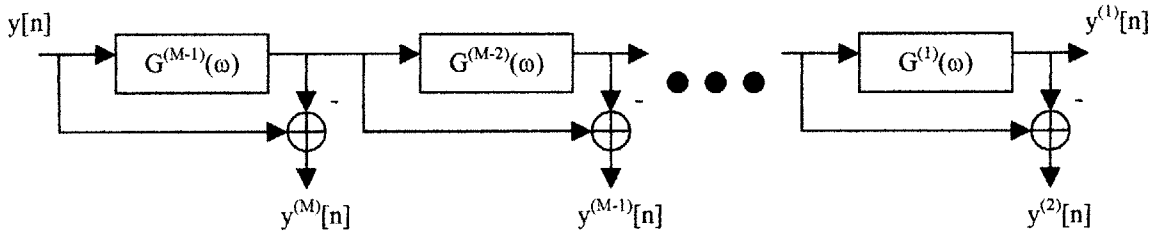


Figure 3.2: M-band decomposition of noisy signal y[n]
by a series of M-1 low-pass filters.

To understand the M-band segmentation process better, take the example of dividing a signal into M = 3 distinct channels. Assume that the input signal is simply an impulse, so that each output channel in Figure 3.2(a) is composed of its corresponding band pass filter frequency response $H^{(i)}(\omega)$. Figure 3.3 presents such a system. Notice how the subtraction of each low-pass filter input and output produces the expected band-pass filters.

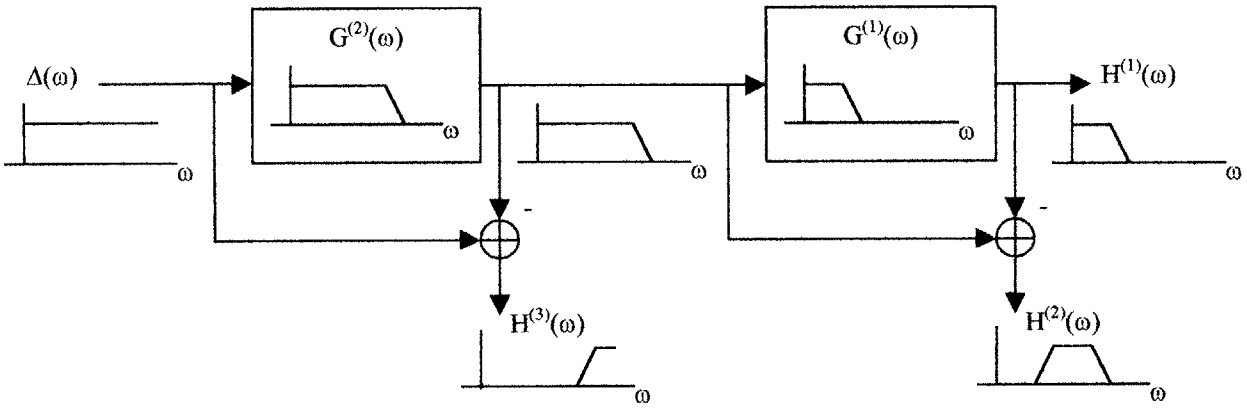


Figure 3.3: Illustrative example of M-band decomposition for M = 3 channels.

3.2.2 Adaptive Analysis Window

After M-band decomposition is achieved, a varying length analysis window is applied to each channel. Traditionally, a fixed length window is used on the whole signal

and the windowed segments are assumed stationary. This assumption is not valid, since stationarity varies for different classes of sounds and different speakers. An adaptive length window is expected to do a better job in analyzing the speech signal.

Since the window's length varies according to the spectral characteristics of speech, a method for quantifying spectral change in the signal is necessary. In this case, spectral change is quantified by the normalized cross-correlation of the smoothed spectra between different time intervals. This similarity measure between the two signal segments $y_{n1}[n] = \{y[n1], \ldots, y[n1 + N - 1]\}$ and $y_{n2}[n] = \{y[n2], \ldots, y[n2 + N - 1]\}$ is denoted by $Q_y(n1,n2)$ and its mathematical expression is given by

$$Q_y(n1,n2) = \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi} F_{n1}(\omega)\cdot F_{n2}(\omega)d\omega}{\max\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} F_{n1}(\omega)^2 d\omega \ , \ \frac{1}{2\pi}\int_{-\pi}^{\pi} F_{n1}(\omega)^2 d\omega\right\}} \tag{3.4}$$

where N is the sample length of the segments and $F_\eta(\omega)$ is defined as

$$F_\eta(\omega) = \left[\frac{L}{2\pi}\int_{\omega-\frac{\pi}{L}}^{\omega+\frac{\pi}{L}} \left|\sum_{n=\eta}^{\eta+N-1} y[n]e^{-j\Omega n}\right|^2 d\Omega\right]^{\frac{1}{2}}. \tag{3.5}$$

Equation 3.5 denotes $F_\eta(\omega)$ as the smoothed or averaged discrete time Fourier transform (DTFT) magnitude of the degraded signal y[n]. The smoothing process is necessary because sudden changes in the spectrum can distort the cross-correlation measurement. The parameter L is a resolution factor that determines the amount of smoothing. It determines the range of frequency samples taken for averaging. Its value varies from 0.005π to 0.05π, depending on the segment's SNR [9].

Figure 3.4 shows the algorithm employed to determine each window's length. At first, a small segment of length $\tilde{N}$ ($\approx$ 10ms) is taken from the noisy channel and its SNR is estimated. This SNR information is used to set the parameters of the adaptive window algorithm, such as the smoothing factor L, the similarity measure thresholds, the number of samples N in each segment, and the amount of overlap between adjacent segments. After these parameters are set, the smoothed spectra $F_{n1}(\omega)$ and $F_{n2}(\omega)$ of two channel sections $y_{n1}[n]$ and $y_{n2}[n]$ are calculated with Equation 3.5. The normalized cross-correlation of these smoothed spectra $Q_y(n1,n2)$ is found following Equation 3.4. If this cross-correlation measurement is higher than the pre-determined thresholds, then the smoothed spectrum $F_{n3}(\omega)$ of an adjacent segment $y_{n3}[n]$ is found and compared to $F_{n1}(\omega)$ by calculating $Q_y(n1,n3)$. This process continues until the cross-correlation measurement goes below the pre-determined thresholds or until the window's length reaches a maximum of 150ms. When this happens, the length of the window is set and the process starts all over again. The iterations continue until the complete time range of the frequency channel is covered.

Once the window lengths are found for each channel, the windows are constructed to satisfy the constraint

$$\sum_m w_m^{(k)}[n] = 1 \qquad \forall n \qquad (3.6)$$

where $w_m^{(k)}[n]$ is the adaptive window of the m^{th} time interval and the k^{th} channel. Meeting this condition ensures that the original noisy signal can be recovered if no enhancement is done. To preserve this constraint, two thresholds for the similarity measure $Q_y(n1,n2)$ are used. The window's magnitude decreases from 1 to 0 in a sinusoidal fashion starting the moment $Q_y(n1,n2)$ goes below the first threshold and until it goes below the second threshold. The following window increases in magnitude during

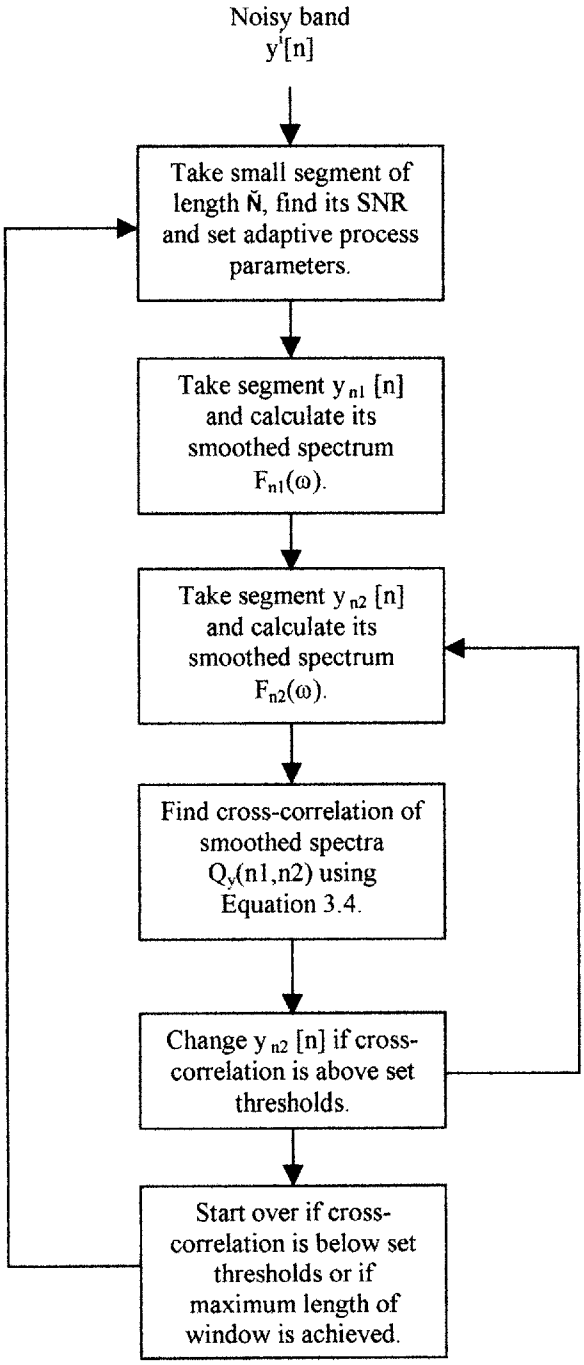


Figure 3.4: Algorithm to determine adaptive window length.

the same time segment. For the k^{th} channel of noisy signal $y^{(k)}[n]$, the m^{th} frame is denoted by $y_m^{(k)}[n]$ and is given by

$$y_m^{(k)}[n] = y^{(k)}[n] \cdot w_m^{(k)}[n]. \tag{3.7}$$

Another step in the time-frequency segmentation process is dividing $y_m^{(k)}[n]$ with three Kaiser windows. The Kaiser windows are used because the cross-correlation based windows have very broad spectra that distort the spectra of the signal segments. Figure 3.5 shows an example. The shapes of the windows suggest that the middle Kaiser window has narrower spectral characteristics than the cross-correlation based window and the Kaiser windows at the edges. It is assumed that the segment selected by the middle Kaiser window will dominate the whole region during overlap add re-assembly. Under this assumption, using the three Kaiser windows will produce better results than using the cross-correlation based windows alone. Notice how the Kaiser windows are designed to add up to one inside the time segment chosen by the cross-correlation window.

3.3 Enhancement of Stationary Regions

After all the stationary time-frequency regions are found, the signal to noise ratio (SNR) of each region is estimated to determine certain parameters of the following enhancement stages. Then an all-pole linear prediction model of the speech signal is used in conjunction with a modified Wiener filter to model and enhance the stationary time-frequency regions.

3.3.1 Local Signal to Noise Ratio Estimation and Parameter Settings

The proposed enhancement system requires SNR estimates of each channel's windowed segment $y_m^{(k)}[n]$, defined here as the Local Signal to Noise Ratio (LSNR) of

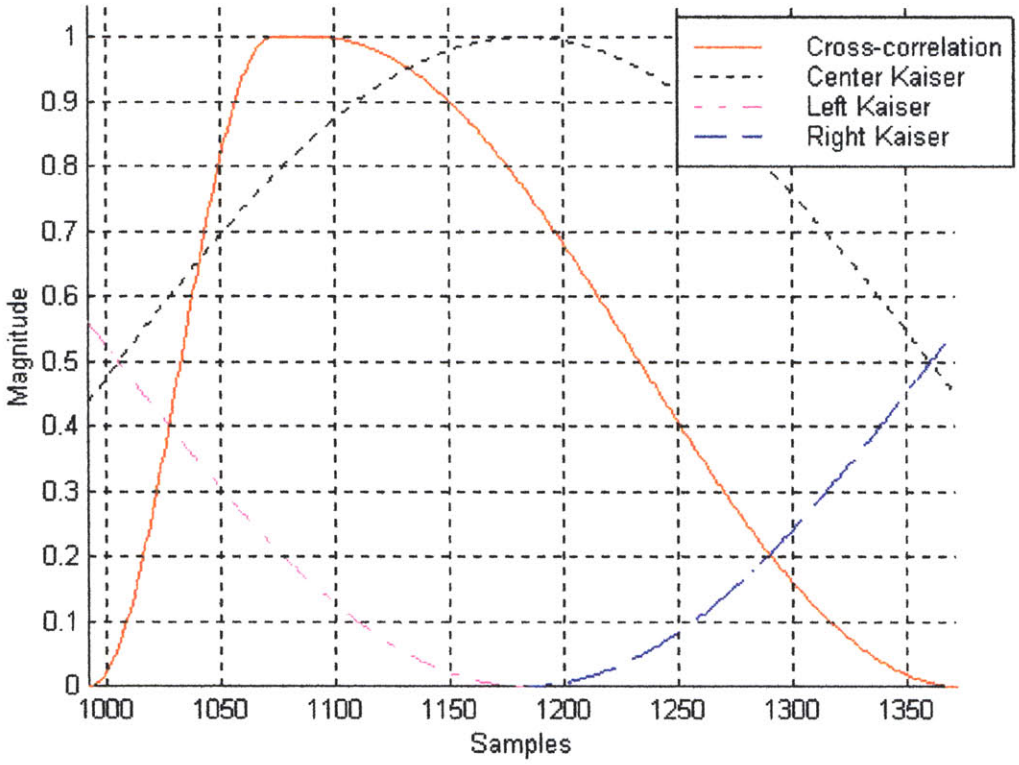


Figure 3.5: Example of three Kaiser windows used to further divide the stationary regions.

the time-frequency region. This LSNR will determine the order of the all-pole model, the modified Wiener Filter parameters, and the number of iterations for the enhancement process.

Using the known band-pass filters $H^{(k)}(\omega)$ composing the filter bank in Figure 3.1 and the window functions $w_m^{(k)}[n]$, it is possible to find an estimate of the LSNR of each region. First, an estimate of the noise spectrum in the k^{th} channel is defined by

$$S_{zz}^{(k)}(\omega) = \left|H^{(k)}(\omega)\right|^2 S_{zz}(\omega) \tag{3.8}$$

where $S_{zz}(\omega)$ is the noise power spectrum prior to band decomposition of the degraded signal.

Now, the LSNR estimate of the region $y_m^{(k)}[n]$ is given by

$$\overline{LSNR}^{(k)}(m) \approx 10\log_{10} \frac{\sum_{n=0}^{N-1} y_m^{(k)}[n]^2 - \frac{1}{2\pi}\int_{-\pi}^{\pi} E\left[\left|Z_m^{(k)}(\omega)\right|^2\right] d\omega}{\frac{1}{2\pi}\int_{-\pi}^{\pi} E\left[\left|Z_m^{(k)}(\omega)\right|^2\right] d\omega} \tag{3.9}$$

where

$$E\left[\left|Z_m^{(k)}(\omega)\right|^2\right] = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{zz}^{(k)} \left|W_m^{(k)}(\Omega-\omega)\right|^2 d\Omega \tag{3.10}$$

and $W_m^{(k)}(\omega)$ is the Fourier Transform of window $w_m^{(k)}[n]$.

Equations 3.8 and 3.10 show how an estimate of the noise energy for each region is found from $S_{zz}(\omega)$, by taking into consideration the effects of the filtering and adaptive windowing processes. On the other hand, Equation 3.9 defines the LSNR estimate of the region $LSNR^{(k)}(m)$ in terms of the defined noise energy and the energy of the degraded speech.

As discussed earlier, the LSNR information is used to set the all-pole model and Wiener filter parameters in the system's enhancement stage. In general, regions with high LSNR are modeled with a higher order all-pole model than low LSNR regions. It is assumed that more original signal is present in high LSNR segments and therefore more poles must be used to model these regions. Furthermore, regions with very low LSNR are "erased" or "smoothed out" from the signal, since any speech information in them is too noisy to be recoverable.

Another criterion used to establish the model order and Wiener filter parameters is the location of low LSNR regions in frequency. If a signal has low LSNR at high frequency channels, it might be necessary to let some of that noise get through during the enhancement stage if it is suspected that the region is part of an unvoiced sound [9]. These special cases are identified by comparing the LSNR of the region in question and the LSNR of the same time segment at the base-band channel, as presented in Figure 3.6. In this way, the system incorporates general speech knowledge of unvoiced sounds into the enhancement process.

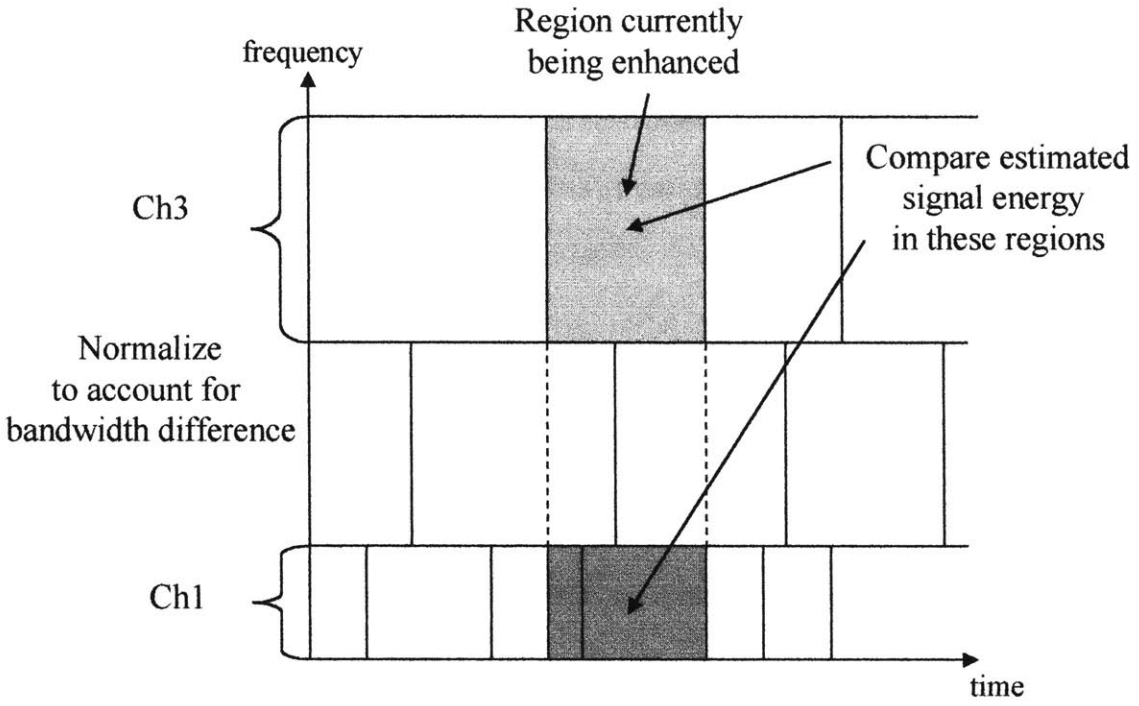


Figure 3.6: Example of LSNR comparison of the same time segment in different channels.

3.3.2 Selective Linear Prediction and Modified Wiener Filter

After selecting the model parameters, selective linear prediction (SLP) is used to calculate the speech model coefficients for each time-frequency region. SLP is necessary

since the degraded signal was segmented into M frequency bands and the linear prediction coefficients inside each filter pass-band are necessary.

Suppose a filter $H^{(i)}(\omega)$ has a pass-band with a frequency range $\omega_1 \leq \omega \leq \omega_2$. To find the all-pole model coefficients with SLP, the spectral region $[\omega_1\ \omega_2]$ gets mapped to $[0\ \pi]$ using the linear conversion proposed by Makhoul [3]

$$\hat{\omega} = \frac{\pi(\omega - \omega_1)}{\omega_2 - \omega_1}. \tag{3.11}$$

The linear prediction coefficients of the region are then calculated following the iterative process in Figure 2.3. The number of iterations needed is determined by the LSNR of the region and it is kept small ($\leq$ 3 iterations) to minimize the model's bias to frequency components of high energy. Afterwards, the stop-band regions are modeled with low order models ($\approx$ 8 poles), and the modeled spectrum is pieced together.

Following the SLP model, a modified Wiener filter is used to enhance the region. The Wiener filter $\Gamma_W^{(k)}(m,\omega)$ for the region in the k^{th} channel and m^{th} time interval is defined by

$$\Gamma_W^{(k)}(m,\omega) = \frac{P_s^{(k)}(m,\omega)}{P_s^{(k)}(m,\omega) + c^{(k)}(m) \cdot P_z^{(k)}(m,\omega)} \tag{3.12}$$

where $P_s^{(k)}(m,\omega)$ and $P_z^{(k)}(m,\omega)$ are the region's SLP model and the noise spectrum respectively. The order of the all-pole model $P_s^{(k)}(m,\omega)$ is varied depending on the region's LSNR and the LSNR measurements found across frequency bands. The parameter $c^{(k)}(m)$ is also varied according to local conditions. After filtering, the enhanced segments are assembled into their corresponding frequency channels by the

overlap add method. Finally, the M channels are added together to form the enhanced signal $\hat{s}[n]$.

3.4 Overall Enhancement System

An overview of the overall system is shown in Figure 3.7. Degraded speech y[n] enters the system and is decomposed into M = 3 frequency channels by an M-band filter bank. The frequency channels have a frequency range of 0 to 1kHz, 1 to 3kHz, and 3 to 5kHz respectively. Figure 3.8 shows time plots of the 3 channels for the speech signal "That shirt seems much too long", sampled at 10kHz and degraded with white noise at 15dB SNR.

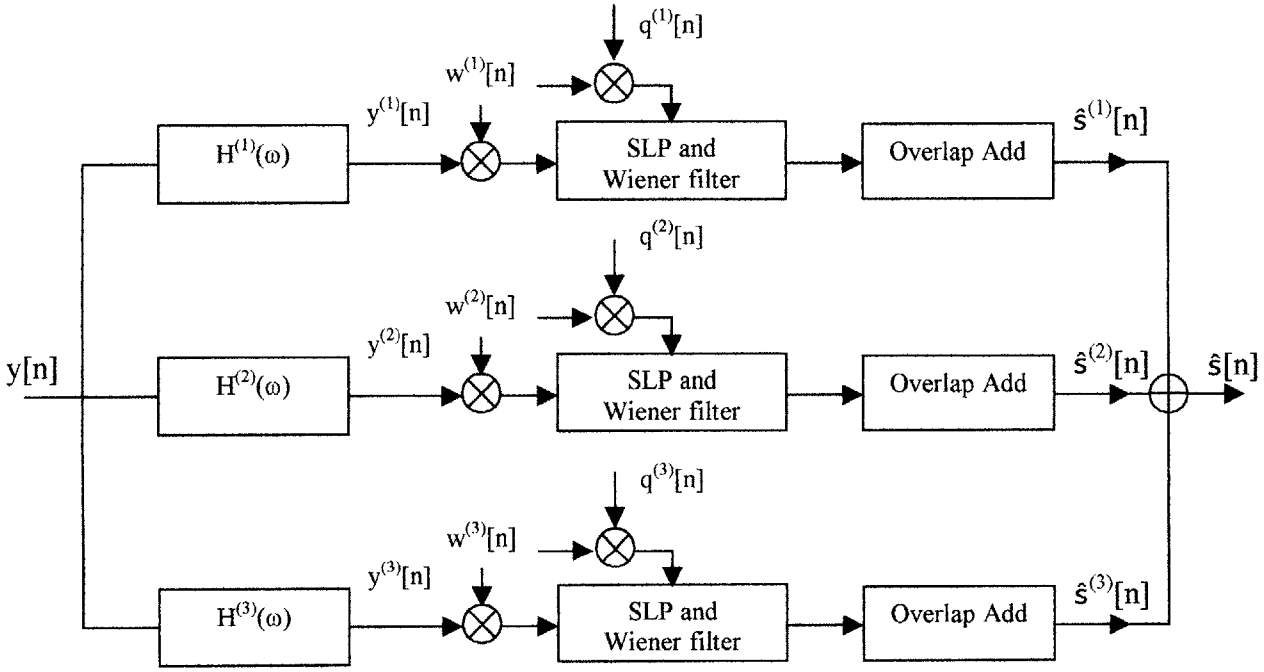


Figure 3.7: Developed enhancement system. The degraded speech signal y[n] is filtered by $H^{(i)}(w)$. Each channel $y^{(i)}[n]$ is then windowed adaptively by $w^{(i)}[n]$ and enhanced by the modified Wiener filters, according to each region's LSNR characteristics. The enhanced speech output is $\hat{s}[n]$.

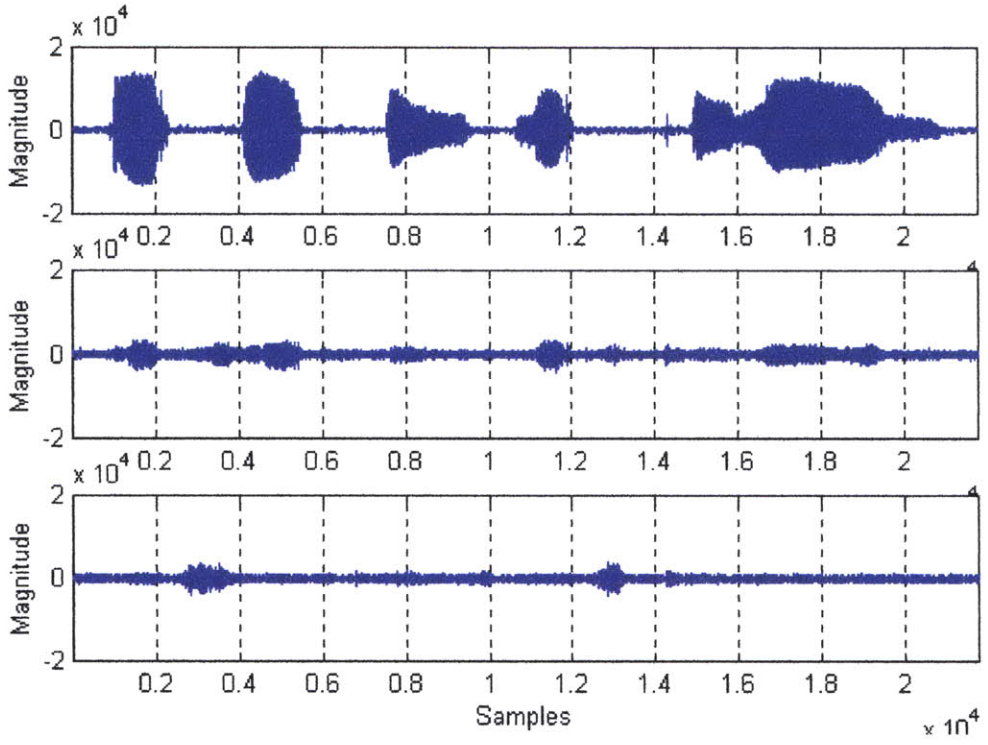
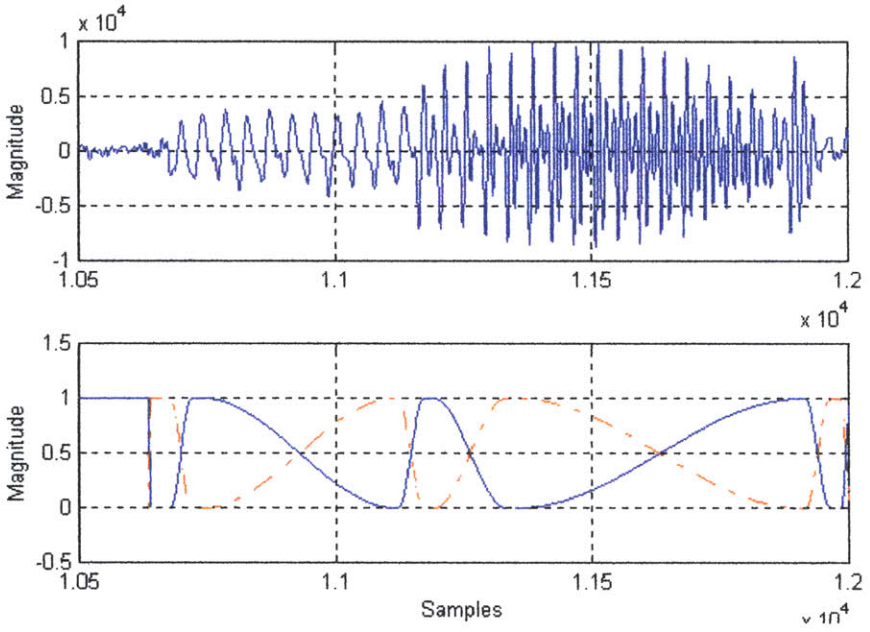
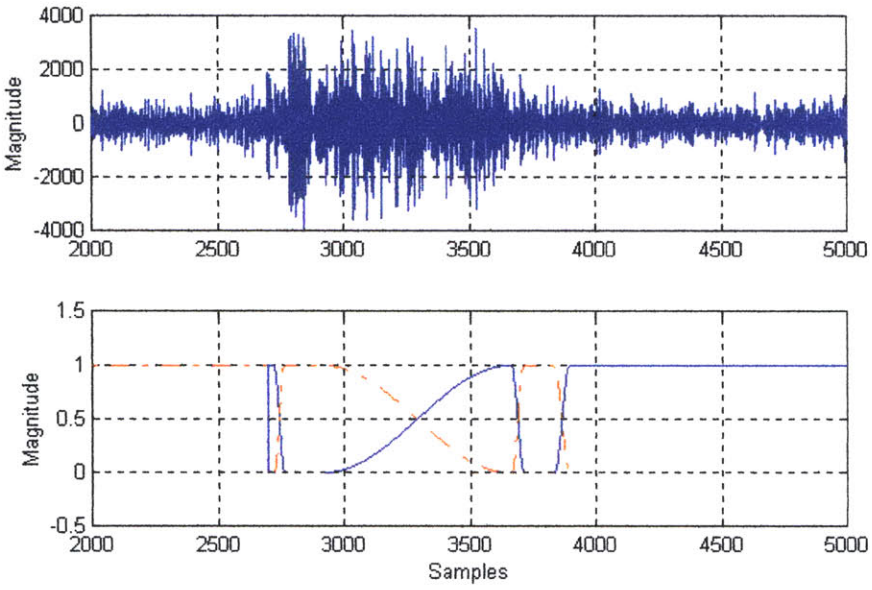


Figure 3.8: Example of speech segmented into three frequency bands. The channels increase in frequency from top to bottom.

Afterwards, the channels get windowed according to their time-varying spectral characteristics. Figures 3.9 (a) and (b) show examples of the cross-correlation based windows used in two of the three channels. The three Kaiser windows mentioned are applied after using the cross-correlation based windows, and only if the region is longer than 10ms.



(a)



(b)

Figure 3.9: Example of adaptive length windows used in the first and third channels of the speech signal "That shirt seems much too long", degraded with additive white noise at an SNR of 15dB.

In the next stage, the SLP and Wiener filter parameters are set using LSNR information from the current region as well as from other frequency bands. The signals $q^{(i)}[n]$ in Figure 3.7 represent extra information from other frequency bands that might be used while the SLP and Wiener filter parameters are being set. Regions with LSNR $\geq$ 30 dB are considered to have high SNR, hence a high number of poles ($\approx$ 50) are used to model them. Meanwhile, regions with LSNR $\leq$ 1 dB are considered to have low LSNR. These regions can be "smoothed out" by the Wiener filter or modeled with a small number of poles ($\approx$ 2), depending on the LSNR at other frequency bands.

After Wiener filtering the regions, each enhanced channel is assembled by overlap-add. The channels are then added to obtain the enhanced speech signal $\hat{s}[n]$. Figures 3.10, 3.11 and 3.12 show spectrograms of the speech signal "That shirt seems much too long." Figure 3.10 is a spectrogram of the clean speech signal without any added noise. Figure 3.11 shows the same signal degraded by additive white noise at an SNR of 15dB. Figure 3.12 is a spectrogram of the enhanced version of the signal, obtained with the algorithm described in this chapter. Notice the similarities and differences between the original and enhanced versions

Objective comparisons as well as informal listening between this and other enhancement methods clearly show that this algorithm is preferable to traditional speech enhancement systems [2, 9]. However, certain aspects of the algorithm could be improved. For example, the adaptive window length could be constrained to have a certain maximum value depending on initial SNR measurements. In addition, the constraints on Wiener filters for high frequency/low LSNR regions could be relaxed to let some of the original speech come through. These and other modifications could be useful in improving the intelligibility of the speech output and increasing the algorithm's computational efficiency.

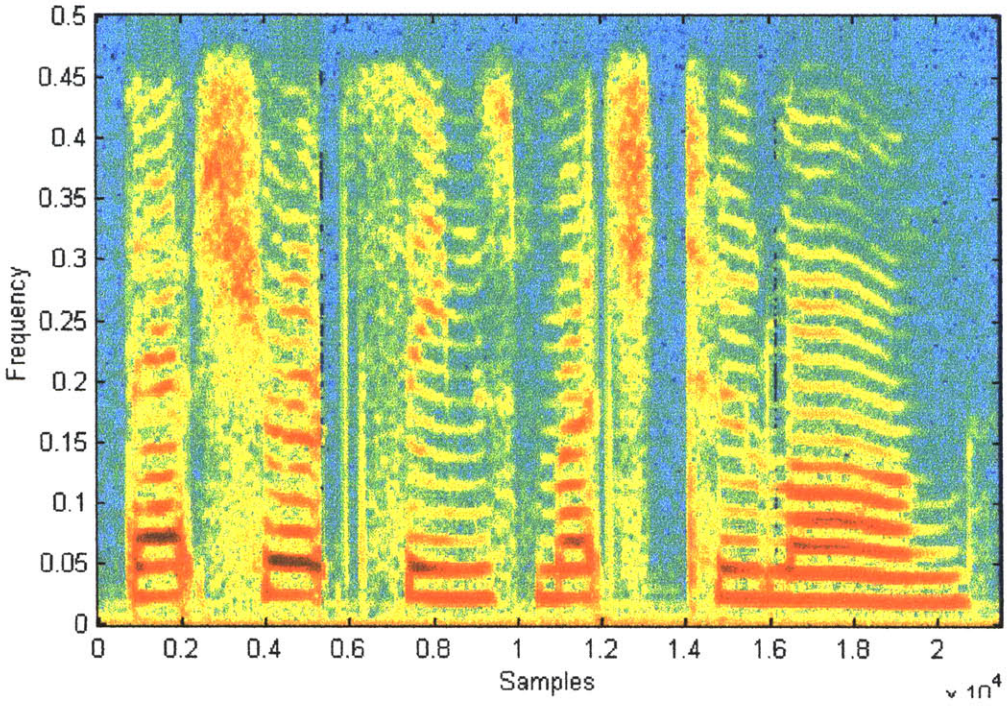


Figure 3.10: Spectrogram of the clean speech signal "That shirt seems much too long."

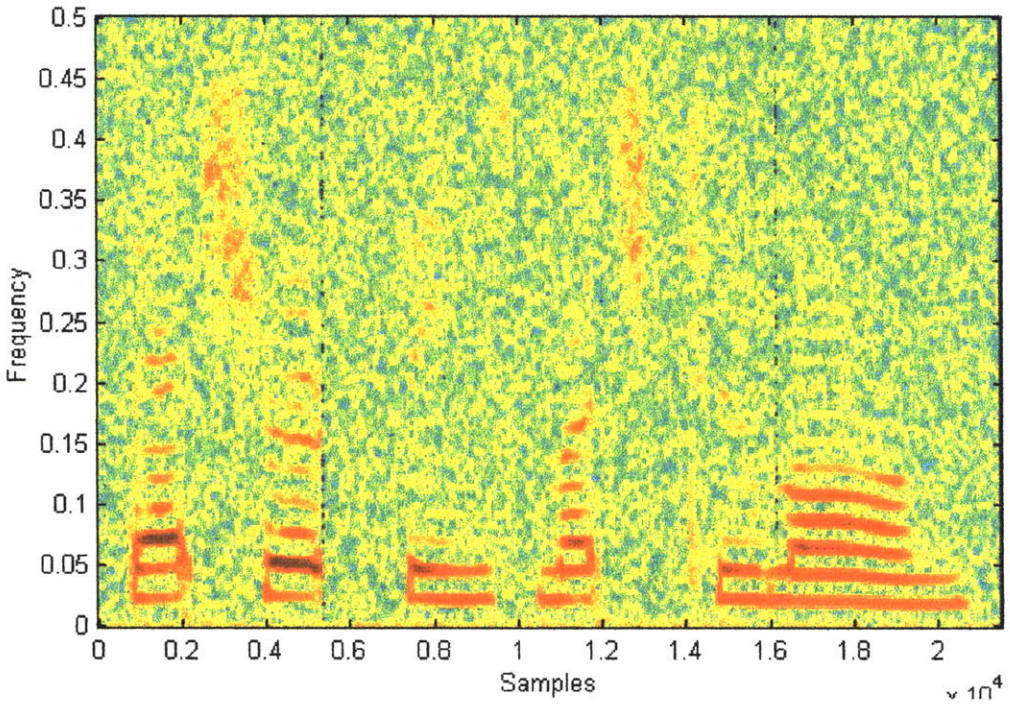


Figure 3.11: Spectrogram of speech signal degraded by additive white noise at 15dB SNR. This is an example a possible input to the system in Figure 3.7 (y[n].)

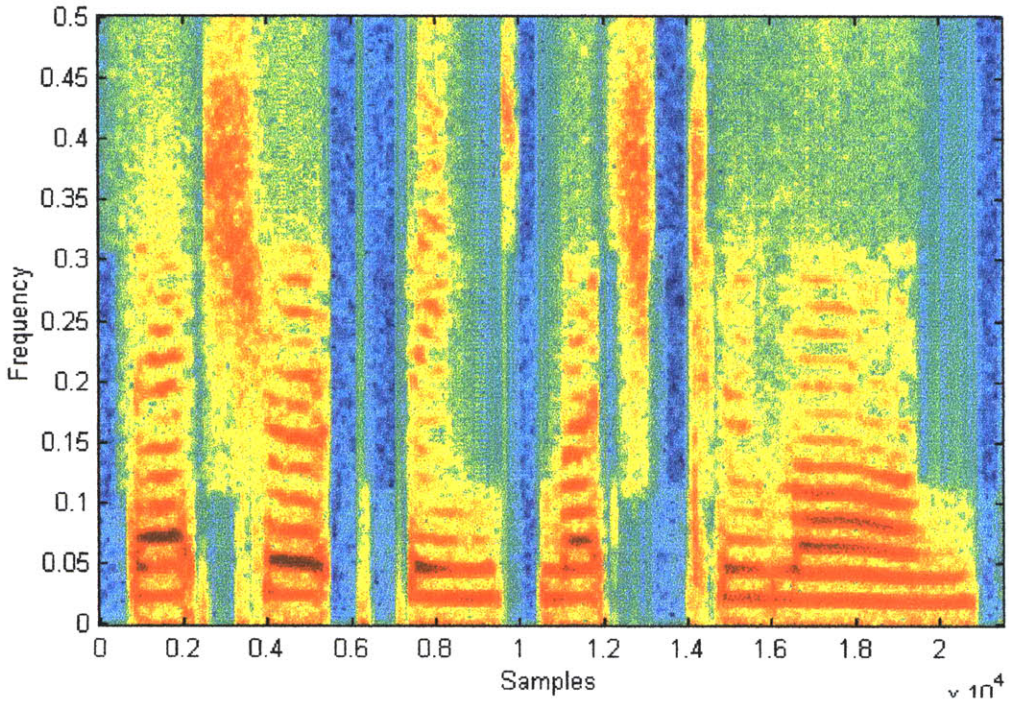


Figure 3.12: Spectrogram of the enhanced speech signal. This is the system's output $\hat{s}[n]$ with the input y[n] in Figure 3.11.

Chapter 4

Modifications to Novel Speech Enhancement System

4.1 Introduction

The enhancement algorithm presented in the last chapter performs better than traditional algorithms. However, it has some disadvantages. For example, the thresholds for the LSNR classification of regions and for estimating the analysis window length must be defined very accurately. Any errors in setting these parameters can cause inaccurate region models and artifacts in the enhanced speech. In addition, since the analysis windows follow cross-correlation measurements, their shape is not very symmetric. This greatly affects the spectral characteristics of the windowed regions and causes the algorithm to make inaccurate LSNR decisions. Applying three Kaiser windows partially solves the problem but it also reduces the computational efficiency of the system. The algorithm also has some difficulty in modeling fricatives and plosives, because the LSNR at the high frequency bursts of these sounds is usually very low and the algorithm tends to smooth them out. These problems cause the system to work better with certain sentences or certain types of speech depending on the specified parameters.

To compensate for some of the system's shortcomings, certain modifications were made to increase its computational efficiency and improve the intelligibility of its output speech. This chapter explains the modifications implemented and the advantages they provide to the original algorithm.

4.2 Alterations to Stationary Region Segmentation Stage

As explained in the previous chapter, the stationary region segmentation of the speech signal is achieved by a two step process. First, the degraded signal goes through a filter bank to separate it into different frequency channels. Then, adaptive length

windows and Kaiser windows are applied to each channel. The length of the analysis windows is found using a cross-correlation based similarity measure.

The modified system presented here maintains the same M-band decomposition set-up as its predecessor. A set of low-pass filters are designed to divide the signal in three channels, with a frequency range of 0 to 1kHz, 1 to 3 kHz and 3 to 5 kHz respectively (assuming a sampling rate of 10kHz.) The filtering scheme is the same as that presented in Figure 3.2.

The only revisions in the stationary region segmentation stage were made in the adaptive length analysis window. Chapter 3 explained how certain parameters of the adaptive windowing process were set by SNR measurements from the first 10ms of signal inside the emerging window. In the modified system, the maximum length of the cross-correlation based window is also changed according to the initial SNR measurement (previously, this length was fixed to 150ms.) This is done because high SNR segments usually have voiced components that are modeled more accurately with longer windows while low SNR regions must be kept short to keep the algorithm from removing small spectral changes (like short plosive bursts.) Following this train of thought, the system was changed so that a low initial SNR measurement sets a window length constraint of 50ms. Also a high initial SNR measurement sets the maximum window length to 150ms and a medium SNR measurement sets a maximum length of 100ms.

Another big change in the adaptive windowing stage was the elimination of the Kaiser windows. The Kaiser windows were introduced in the system because the spectral characteristics of the cross-correlation based windows were undesirable and caused the system to make wrong LSNR decisions. To eliminate them, the similarity thresholds between the spectra of small segments were separated more to produce smother changes in the analysis window's shape. The elimination of the three Kaiser windows increased the system's computational efficiency significantly, since only one region model per cross-correlation window was needed for enhancement instead of three.

4.3 Modifications to Selective Linear Prediction Modeling and Wiener Filtering Stages

The enhancement system in Chapter 3 finds an all-pole model of the signal inside a channel specified frequency range. However, the modeling is performed using the spectrum of the windowed noisy signal $y_m[n] = y[n] \cdot w_m^{(k)}[n]$ instead of the filtered and windowed signal $y^{(k)}_m[n] = y^{(k)}[n] \cdot w_m^{(k)}[n]$. This hinders computational efficiency, since the noisy signal must be windowed and modeled. To simplify the algorithm, the speech signal was modeled using only the filtered and windowed signal $y^{(k)}_m[n]$. The frequency mapping in Equation 3.4 and the corresponding frequency range for each band-pass filter were used in the selective linear prediction process.

Another improvement in computational efficiency was achieved by disregarding the region models of each channel's stop band segment. Since the band-pass filters attenuate the amount of energy in these regions, it is not necessary to use any poles to model them. In addition, the Wiener filtering stage was integrated into the parameter estimation process for the linear prediction model. Since only the filtered and windowed signal $y^{(k)}_m[n]$ was used in the parameter estimation, there was no need for an extra Wiener filtering stage after the SLP modeling iterations. Figure 4.1 shows how the SLP model and the Wiener filter stages were merged into one process.

The parameters for the Wiener filters were determined according to the LSNR of the regions. The basic guidelines for setting these parameters followed two main ideas: using noise masking at lower frequency regions and maintaining some of the noise at high frequency regions. Since speech signals usually have high energy voiced components at low frequency regions, there is no need to remove too much noise from these regions, since the high energy voiced components will make the residual noise perceptually undetectable (noise masking). At higher frequencies, the bursts of plosives and fricatives must stand out from the rest of the noise, but the other speech segments cannot be completely "erased" because there is a small amount of voiced energy at these higher frequencies. Eliminating these areas makes the output speech sound muffled.

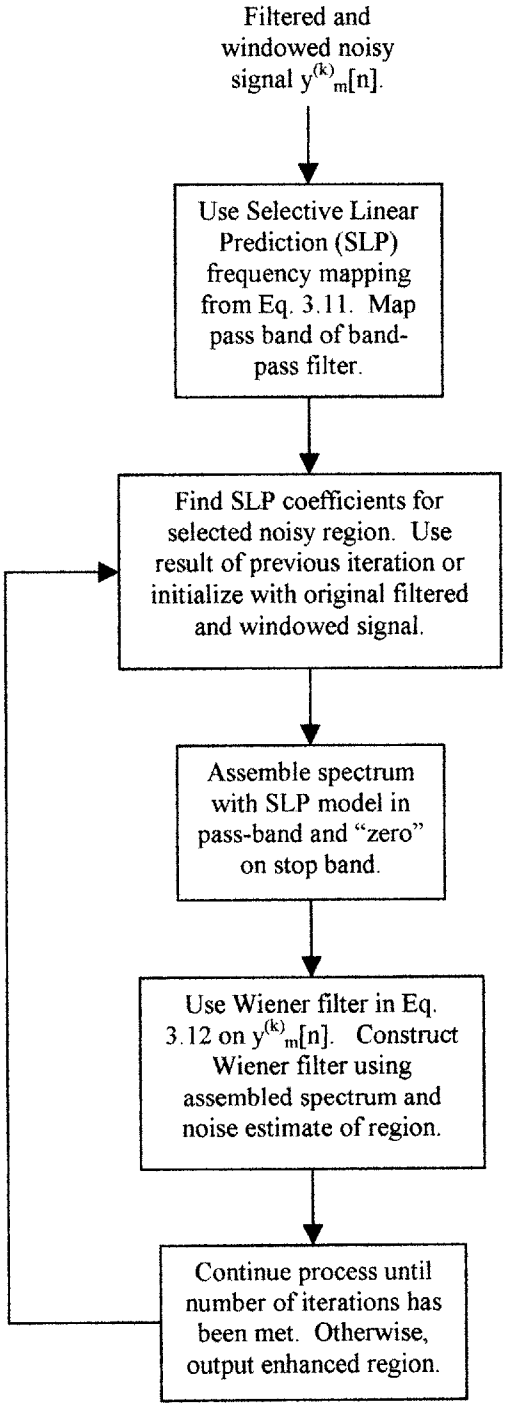


Figure 4.1: Algorithm for finding SLP coefficients and Wiener filtering of noisy regions.

Therefore, it is desirable to relax the Wiener filter constraints on low LSNR regions at higher frequencies (as compared to previous incarnations of the algorithm that removed noise more aggressively in these regions). Following these ideas, the Wiener filter and model parameters for the proposed enhancement system were set as follows:

1) At high frequency regions with low LSNR, low order models were used and noise was removed moderately.
2) At high frequency regions with high LSNR, high order models were used and noise was removed heavily.
3) At low frequencies regions with high LSNR, high order models were used and noise was removed moderately.
4) At low frequency regions with low LSNR, low order models were used and noise was removed heavily.
5) At regions with medium LSNR, parameters tend to vary according to extra LSNR information from regions within the same time segment but at lower frequency bands.

The modeling stage of the system was also altered in the thresholds that separate low, medium and high LSNR regions. In general, regions at higher frequency bands now have thresholds farther apart than regions at lower frequency bands. This scheme was used because at lower frequency bands speech tends to have more voiced components and higher SNR measurements, therefore sharper decisions between low or high LSNR regions can be made. On the other hand, deciding between low or high LSNR regions at high frequencies is harder due to the appearance of high frequency bursts from plosives and fricatives. To help in the decision process, more high frequency regions are classified as having medium LSNR so that information from lower frequency bands is considered. This extra information is instrumental in making decisions on model and Wiener filter parameters for high frequency regions. Figure 4.2 shows the general trends of LSNR classification for a signal separated in three frequency channels.

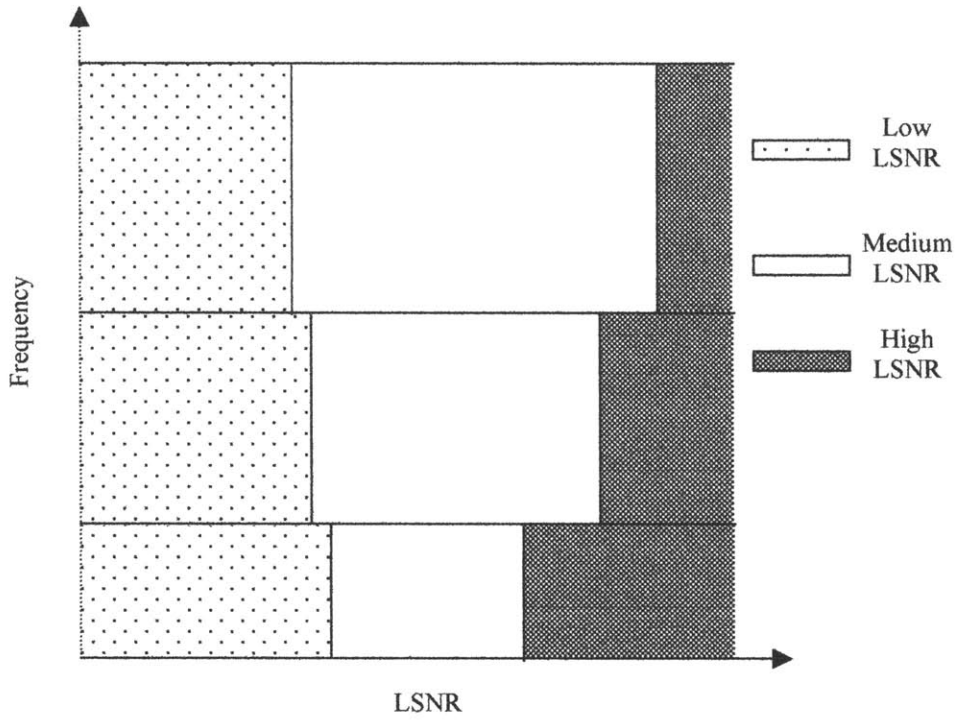


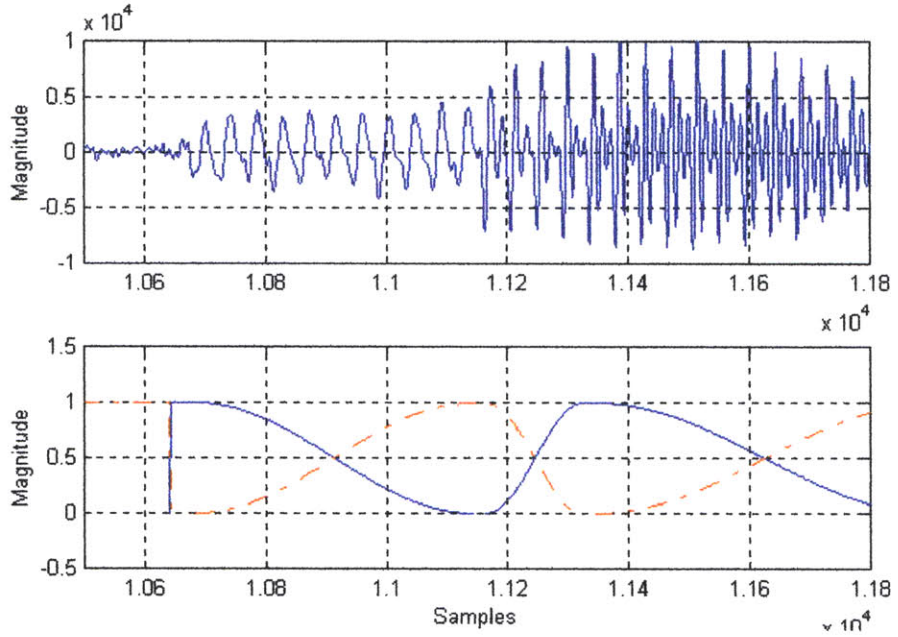
Figure 4.2: LSNR classification trends according to frequency channel.

4.4 Proposed System

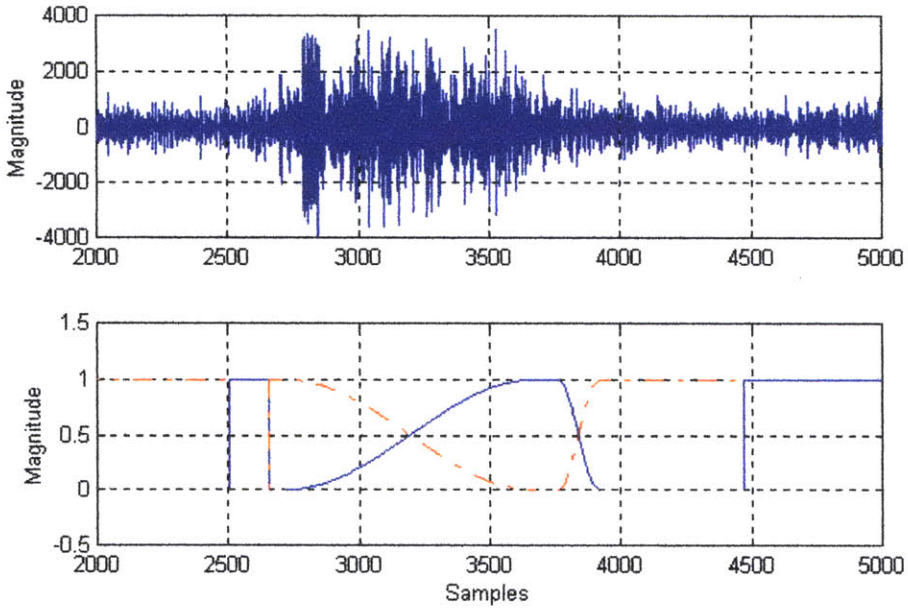
Table 4.1 shows the algorithm's settings and parameters for the adaptive length window process. As explained above, this is the only step from the stationary region segmentation process that's different from the algorithm in Chapter 3. Notice how all the parameters vary according to initial SNR measurements. Notice also how there are two cross-correlation thresholds. Once the similarity measure goes below the high threshold mark, the window starts to decrease in a sinusoidal fashion until it goes below the low cross-correlation threshold. Recall that the window length will be determined either by the cross-correlation thresholds or by the maximum length established. These parameters are kept constant for all channels. Figures 4.3 (a) and (b) show examples of the windows obtained with the parameter settings in Table 4.1.

	SNR > 30dB	SNR < 1dB	1dB ≤ SNR ≤ 30dB
High cross-correlation threshold	1.0	1.0	0.75
Low cross-correlation threshold	0.6	0.1	0.25
Segment length	15ms	10ms	15ms
Segment overlap	2ms	5ms	2ms
Frequency resolution factor	0.015π	0.005π	0.01π
Maximum window length	150ms	50ms	100ms

Table 4.1: Parameters for window length specification.



(a)



(b)

Figure 4.3: Example of adaptive windows of modified algorithm used in the first and third channels of the speech signal "That shirt seems much too long." The signal was degraded with additive white noise at a SNR of 15dB.

The following tables show the chosen set of parameters for the SLP modeling and Wiener filtering stages of the enhancement system. Each of the three bands has its own set of linear prediction and Wiener filter parameters. Notice how the LSNR thresholds change according to the frequency band.

	LSNR > 30dB (High LSNR region)	5dB≤ LSNR ≤ 30dB (Mid LSNR region)	LSNR < 5dB (Low LSNR region)
Model order	50	50	0
Number of iterations	3	2	3
Wiener filter weighting factor	1.0	1.0	1.0

Table 4.2: Parameters for modeling and enhancement of regions in the first frequency band.

	LSNR > 35dB (High LSNR region)	1.5dB≤ LSNR ≤ 35dB (Mid LSNR region)	LSNR < 1.5dB (Low LSNR region)
Model order	50	50	0
Number of iterations	3	2	3
Wiener filter weighting factor	1.25	1.0	0.85

Table 4.3: Parameters for modeling and enhancement of regions in the second frequency band.

	LSNR > 40dB (High LSNR region)	0.5dB≤ LSNR ≤ 40dB (Mid LSNR region)	LSNR < 0.5dB (Low LSNR region)
Model order	50	50	0
Number of iterations	3	2	3
Wiener filter weighting factor	1.5	1.0	0.75

Table 4.4: Parameters for modeling and enhancement of regions in the third frequency band.

The model parameter values presented in these tables are not exclusive. For regions with medium or low LSNR, the model order, Wiener filter, and number of iterations for enhancement vary according to the LSNR of the same time region at different bands. Figure 4.4 shows how the model and Wiener filter parameters are assigned in these special cases. Notice how the modified algorithm uses information from the base band channel as well as from previous lower frequency channels, depending on the LSNR of the current region. In the previous algorithm, only the information from the first band (base band channel) was used.

The next figures show an example of the results obtained with the modified enhancement algorithm. Figure 4.5 is a spectrogram of the clean speech signal "That shirt seems much too long" sampled at 10kHz. Figure 4.6 shows the same signal degraded by additive white noise at an SNR of 15dB. Figure 4.7 is a spectrogram of the enhanced version of the signal, obtained with the modified algorithm. Again, notice the similarities and differences between the original and enhanced version. Also, compare the spectrogram in Figure 4.7 with the spectrogram of the enhanced speech signal shown in Figure 3.12. Note that the new enhancement algorithm produces a smoother speech signal, with less abrupt transitions between spoken and silent segments.

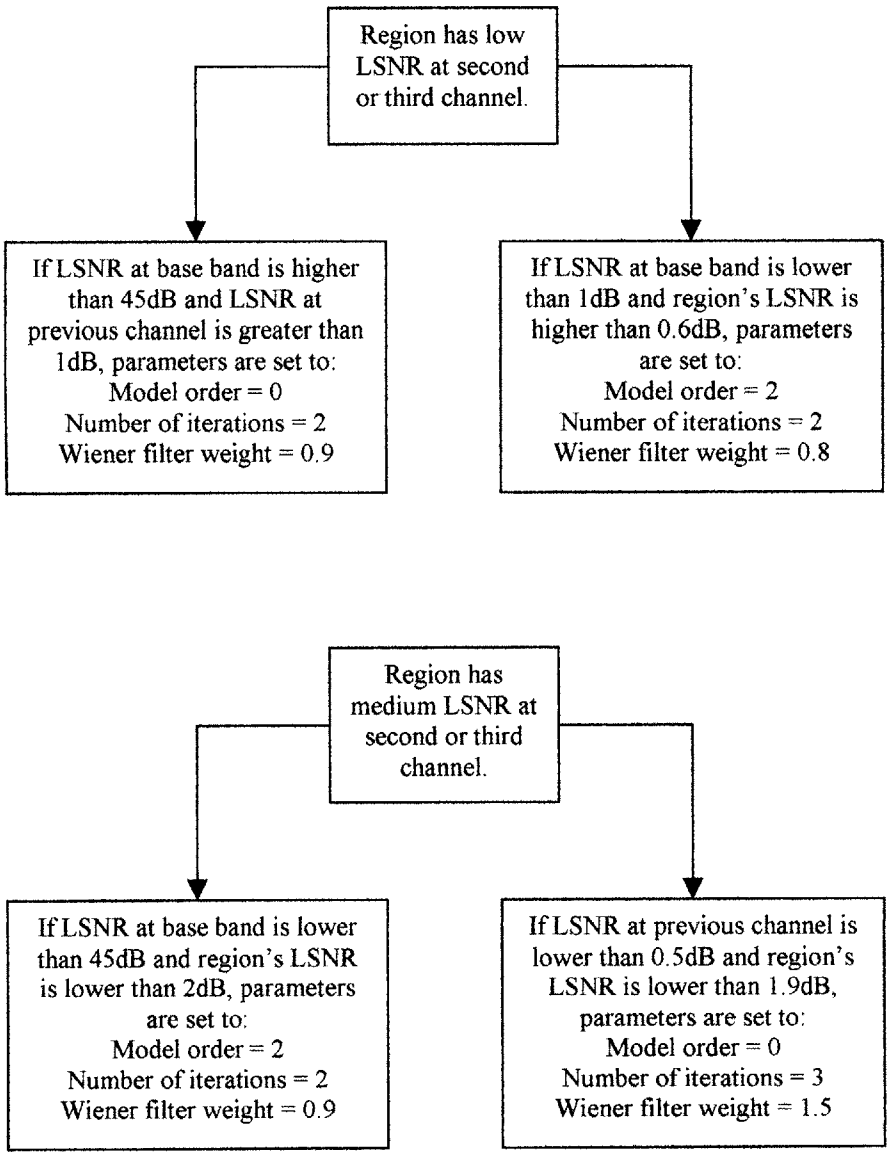


Figure 4.4: Diagram for parameter settings in low and medium LSNR regions using LSNR information from other bands.

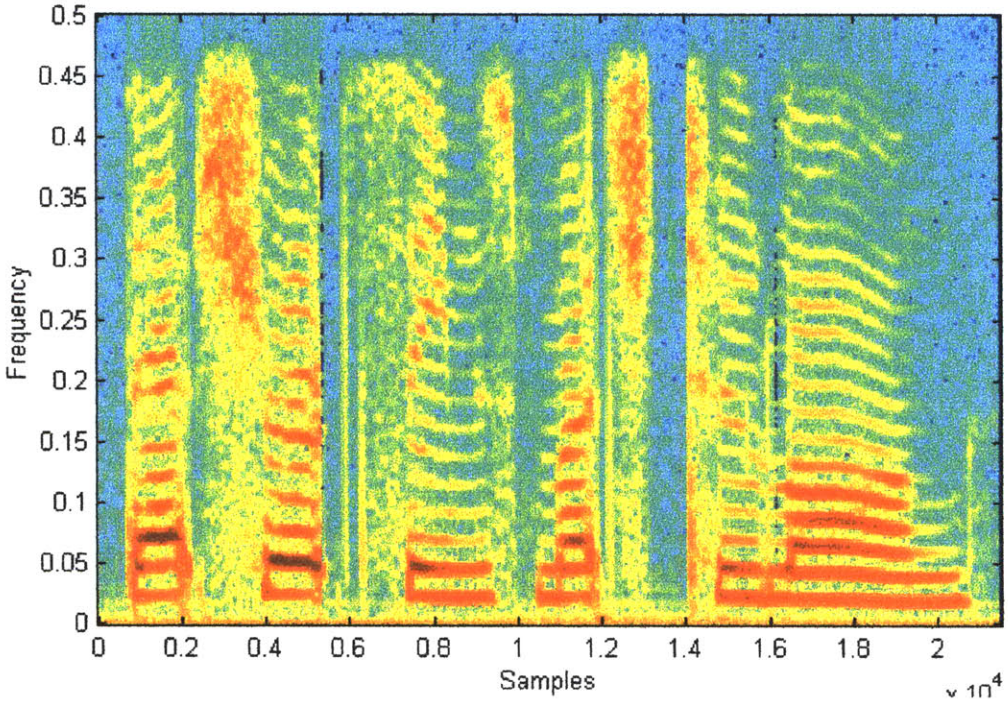


Figure 4.5: Spectrogram of the clean speech signal "That shirt seems much too long."

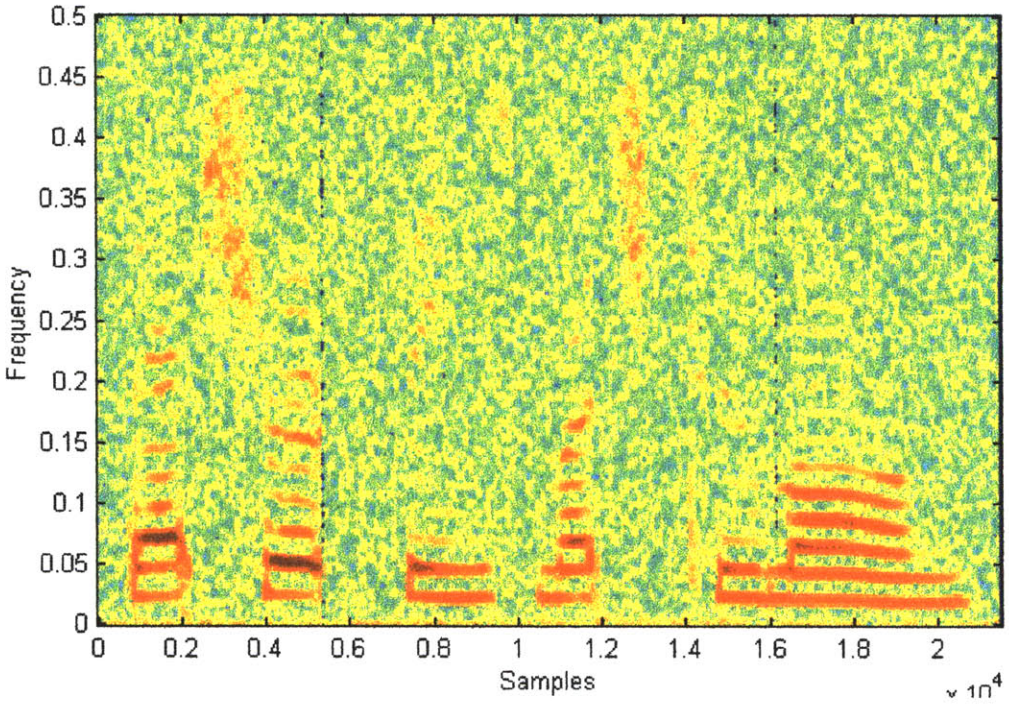


Figure 4.6: Spectrogram of speech signal degraded by additive white noise at 15dB SNR.

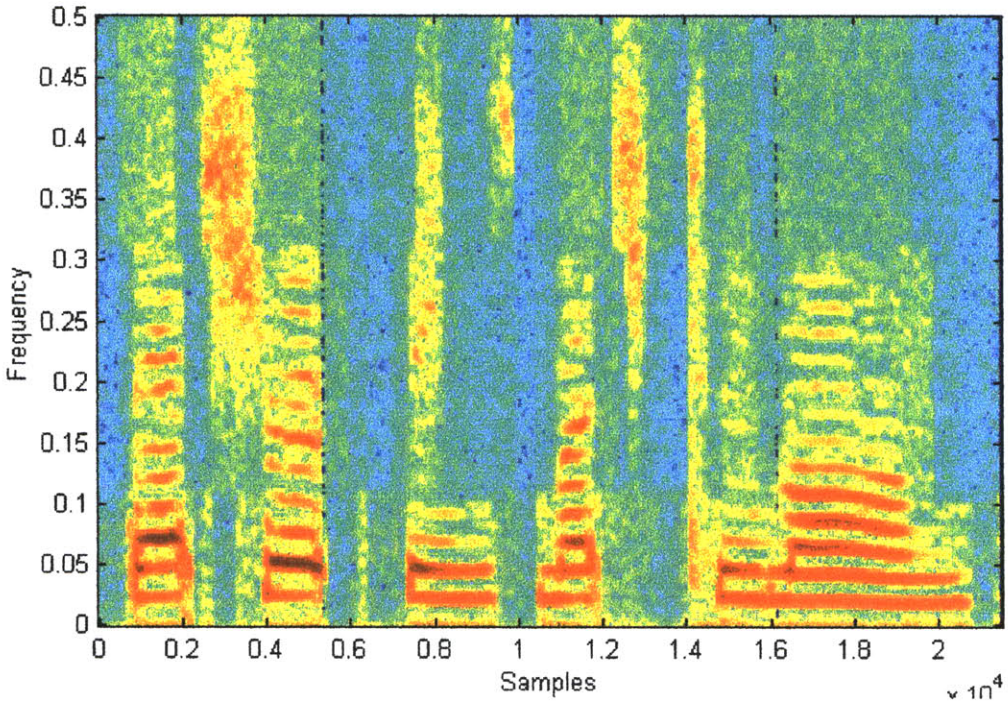


Figure 4.7: Spectrogram of the enhanced speech signal.

Chapter 5

System Evaluation

5.1 Introduction

In the previous chapter, a set of modifications for the speech enhancement system presented in Chapter 3 were proposed. The original system used a filter bank and a set of adaptive length windows to segment the speech signal into nearly stationary time-frequency regions before enhancement. The modifications suggested in Chapter 4 were geared towards simplifying the algorithm for increased computational efficiency as well as achieving a better compromise between quality and intelligibility. These modifications were made mostly in the adaptive windowing and modeling stages of the process.

This chapter discusses the performance of the latest system. It will focus on objective and subjective measurements to quantify improvements in quality and intelligibility. The system will be compared to its previous incarnation, as well as to traditional enhancement techniques.

5.2 Proposed Method Vs Traditional Methods

The proposed method will be compared in terms of quality, intelligibility and computational efficiency to previously developed enhancement systems. The quality and intelligibility will be evaluated in terms of objective measurements, such as segmental SNR and the Itakura-Saito distance and subjective measurements, mainly the opinion of experienced listeners. Comments on computational efficiency gains are also presented

5.2.1 Objective Measures

The two main objective measures used in this thesis are segmental signal-to-noise ratio (SNR) and the Itakura-Saito measure. Segmental SNR is the average SNR over short segments of speech waveform. It is considered a good estimator of speech quality because the segmentation in the SNR computation permits the measure to assign equal weight to all portions of the speech signal [2]. However, this measure is not flawless. If the speech has intervals of silence, any amount of noise will give rise to a large negative SNR for those regions that could bias the overall measure. To solve this problem, the silent frames could be identified and excluded from the computation or the SNR estimate can be bounded with an arbitrary threshold (e.g., -5dB < sSNR < 25dB.)

The Itakura-Saito measure can be used to quantify dissimilarities between all-pole models of the original and enhanced speech. Since the human auditory system is relatively insensitive to phase distortion, many enhancement systems focus only on the magnitude of the speech spectrum. Measures based on SNR do not provide a meaningful measure of performance when the two waveforms differ in their phase spectra because they obtain a distortion measure based on sample by sample differences in the original and processed time-waveforms. The Itakura-Saito measure is sensitive only to variations in the magnitude spectrum. Therefore, it is impartial to phase differences between the original and enhanced speech signals. The distance measure is computed between sets of Linear Prediction (LP) parameters estimated over synchronous frames in the original and processed speech.

Figures 5.1 and 5.2 show plots of segmental SNR and Itakura-Saito measures for the enhancement systems presented in Chapters 3 and 4. These results are compared to the measurements obtained with other traditional speech enhancement systems. Notice that the objective measures show the two novel enhancement systems performing better than the traditional systems. In addition, the modified enhancement system of Chapter 4 shows a small improvement in both plots as compared to its previous implementation.

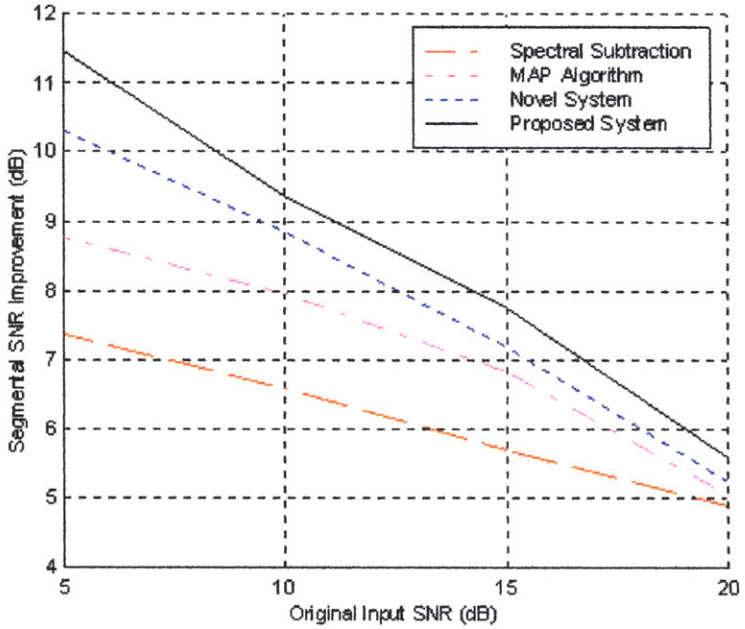


Figure 5.1: Segmental SNR measures for traditional and novel speech enhancement systems.

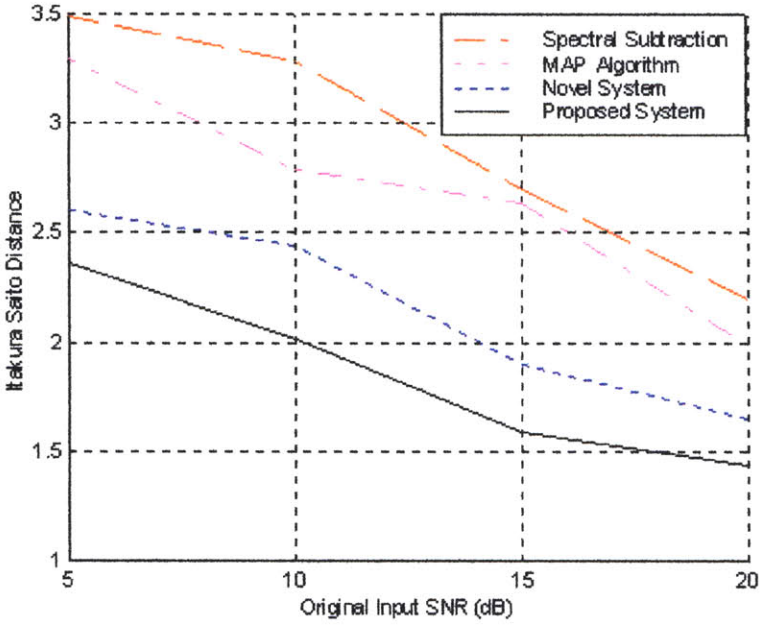


Figure 5.2: Itakura-Saito measures for traditional and novel speech enhancement systems.

5.2.2 Subjective Measure and Computational Efficiency

The main subjective measure used to evaluate the system was informal listening. A small group of speech experts ($\approx$ 5) listened to speech signals degraded at SNR's between 10 and 20 dB. In all the cases, it was agreed that the proposed system produced speech that was more intelligible than the speech produced by the rest of the presented enhancement algorithms.

It is important to point out the big computational improvement that the proposed enhancement algorithm brings to the process. Simplifying the original algorithm (presented in Chapter 3) by removing various steps in its windowing and modeling stages reduced processing by a factor of 10. For example, the previous enhancement system needed about 4.3 minutes to process 21 seconds of speech. The modified algorithm from Chapter 4 needs 28 seconds to process the same signal in the same computer environment. The systems were implemented in C, using Ultra 5 Sun Workstations with a SunOS 5.6 operating system. Further advancements could make the system work at real time speeds.

Chapter 6

Summary and Future Research

6.1 Summary

This thesis considered the problem of reducing noise in degraded speech. It was desired to develop a system that would maximize noise reduction while minimizing speech distortion, always taking into consideration that noise reduction often leads to speech distortion. To achieve this goal, a balanced tradeoff between these opposing goals was targeted by exploiting local characteristics of stationary time-frequency regions. This technique is very different from traditional enhancement techniques that try to achieve this tradeoff over the entire fixed-length windowed speech segment.

The systems presented in this thesis exploit both time and frequency localized properties of speech. Local characteristics are obtained from stationary regions selected by decomposing the signal into different frequency bands and applying adaptive length windows to each channel. The enhanced spectrum of each stationary region is estimated with an all-pole model using Selective Linear Prediction (SLP), which allows the process to model only the spectral region of interest. By modeling the local spectrum, either independently or dependently of other time-frequency regions, and adjusting the model and Wiener filter parameters according to each region's local signal to noise ratio, a balanced tradeoff between noise reduction and speech distortion was achieved. The proposed system does not suffer from tonal artifacts like spectral subtraction or from bias problems like the MAP algorithm. Modifications were made in several stages of the algorithm to improve its computational efficiency and the intelligibility of the enhanced speech.

Results based on informal listening and objective measures such as segmental SNR and Itakura-Saito distance indicate that the modified enhancement system

performed better than traditional algorithms. The results also showed that the modified system radically increased the computational efficiency of its previous incarnation and improved to a lesser extent the quality of the output speech.

6.2 Future Research

The ideas presented as part of this new speech enhancement technique have considerable potential for further research. For example, the time-frequency regions could be identified using time-frequency representations or techniques other than M-band segmentation and adaptive windowing. A few possibilities are the Wavelet transform, the Garbor transform and the Wigner distribution. In addition, noise reduction in each time-frequency segment could be implemented using a variety of different enhancement techniques that may or may not include speech models, such as those based on noise masking properties of speech. Finally, several other local characteristics of the time-frequency regions can be used to determine model and Wiener filter parameters such as the entropy or voicing state of the regions. More study of these and other alternatives is necessary to further improve the described system.

References

[1] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York, New York: McMillan, 1993.

[2] C. D. Yoo, *Speech Enhancement: Identification and Modeling of Stationary Time-Frequency Regions*. Ph. D. thesis, June 1996.

[3] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561-580, April 1975.

[4] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-26, pp. 197-210, June 1978.

[5] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. New Jersey: Prentice Hall, 1978.

[6] H. Drucker, "Speech processing in high ambient noise environment," *IEEE Trans. Audio Electro-acoustics*, vol. AU-16, pp.165-168, June 1968.

[7] R. J. McAulay and M. L. Malpass, "Speech enhancement using soft-decision noise suppression filter," *IEEE Trans. On Acoustics, Speech and Signal Processing*, vol. ASSP-28, pp. 137-145, April 1980.

[8] D. W. Griffin and J. S. Lim, "A new pitch estimation algorithm," *Int. Conf. On Digital Signal Processing*, September 5-8, 1984.

[9] C. D. Yoo and J. S. Lim, "Modeling Stationary Time-Frequency Regions of Noisy Speech," Korea Telecom, June 1997.