uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Wei Zhang
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Electrical Engineering)
GRADE / DEGREE

School of Information Technology and Engineering
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Speech Enhancement Based on Perceptual and Statistical Models of Speech
TITRE DE LA THÈSE / TITLE OF THESIS

Tyseer Aboulnasr
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Peter Kaball

Rafik Goubran

Richard Danserean (Carleton University)

Martin Bouchard

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

SPEECH ENHANCEMENT BASED ON PERCEPTUAL LOUDNESS AND STATISTICAL MODELS OF SPEECH

by

WEI ZHANG

Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies,

University of Ottawa

In partial fulfillment of the requirements

For the degree of Doctor of Philosophy

in Electrical Engineering

School of Information Technology and Engineering

Faculty of Engineering

University of Ottawa

Your file Votre référence
ISBN: 978-0-494-61400-6
Our file Notre référence
ISBN: 978-0-494-61400-6

Abstract

This dissertation is concerned with speech enhancement based on the statistical and loudness models. We will study the field of speech enhancement with the objective of improving the quality of speech signals in noisy environments.

First, speech enhancement based on the Laplacian model for speech signals is reviewed. The performance is shown to be limited by the accuracy of the Laplacian parameter estimation in the noisy environment. A recursive version is proposed to estimate the Laplacian model parameters using the enhanced speech and then use these estimated parameters to re-enhance the original noisy speech again. This approach achieves better parameter estimation and hence further improvements of speech quality.

Next, loudness models for speech are reviewed. Considering that it describes the human hearing system better than the spectrum, the fundamental approaches of spectral subtraction are extended to the loudness domain. We propose the loudness subtraction approach. The tests are done for subtraction with different α values in the loudness model. Simulations show that the quality of enhanced speech can be optimized by choosing the appropriate α for a given input SNR. Thus, an adaptive-α subtraction model is proposed. The simulations show it can further improve the performance of spectral subtraction.

Then, the proposed loudness subtraction with fixed α is shown to provide better results overall than the classical spectral subtraction, even though noise residue and unpleasant

artifacts are still high in the enhanced signal. Loudness over-subtraction is then proposed to further reduce these artifacts/noise. Extensive simulation studies are conducted showing clear improvement over other subtraction type approaches.

Finally, we proposed a Maximum Likelihood-based (ML) speech enhancement algorithm in the loudness domain. It is an optimal speech enhancement algorithm based on the ML criteria in the loudness domain, given the loudness of the noisy speech and the noise estimate. The Laplacian model and the Gaussian model of speech are used separately for comparison. Both approaches shows significant improvement of quality. It is shown that the Laplacian model leads to better preservation of the speech and the Gaussian model leads to better noise reduction.

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude towards my supervisor Dr. Tyseer Aboulnasr for giving me this opportunity to work with her. Her knowledge, experience, common sense and perspectiveness had help me throughout this thesis.

I also want to thank my parents whose support over the years has made this work possible. Also thanks to the help from all my colleagues in the SPOT research group.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Thesis Motivation

Since the 1960s, speech processing has been an active area of research in digital signal processing [8]. In the last several decades, speech processing has had direct applications in society. A variety of areas have been discussed by many researchers, like speech coding [29] [52], speech recognition [89], speech enhancement [40] [80] and so on. A comprehensive review of speech research can be found in [70]

Over the past four decades within speech processing, the enhancement of speech degraded by additive background noise has received considerable attention. The main goal of speech enhancement is to improve the performance of speech communication in a noisy environment. Depending on the application, the objective of the enhancement system may be to improve the overall quality, increase intelligibility, reduce listener fatigue and so on. Some systems may meet several of these objectives at the same time. The recent research of speech enhancement are reviewed in [38].

In most speech enhancement algorithms, it is assumed that the speech is degraded by

additive noise which does not depend on the clean speech. Some other practical noise sources can be transformed into additive noise [80]. For example, a multiplicative or convolutional noise degradation is converted to an additive noise degradation by a homomorphic transformation [3].

Many speech enhancement systems [49] [71] [80] [87] [94] [102] improve the signal to noise ratio (SNR) as a way to provide higher quality. However, almost all of these systems reduce the intelligibility, which may be acceptable to listeners, particularly when the test material is familiar. On the other hand, Niederjohn [92] claimed that certain process of filtering can improve intelligibility while quality may be degraded.

The auditory model of speech signals has been used extensively in speech enhancement. The human hearing system has been thoroughly studied and many speech enhancement algorithms are designed to be more suitable for subjective listening. The masking property of human hearing system has lead to many developments in speech enhancement [73] [74] [109], allowing for actual improvement in perceived quality as compared to the mathematical improvement in SNR.

Even though loudness of speech has been considered as one of the most important measures to simulate the human hearing system, few researchers have proposed the speech enhancement algorithms in the loudness domain. In [93], the estimated loudness of the noise is subtracted from the loudness of the noisy signal to enhance speech. This algorithm performs well when the input speech SNR is low, but it leads to large distortion when the SNR is high. This is because the loudness of the noisy signal is not simply the sum of the loudness of noise and clean signal. In this thesis, the loudness model will be discussed thoroughly and a new loudness subtraction algorithm is proposed.

Statistical modelling of the speech is discussed thoroughly in [118]. A Laplacian model

is found to be more appropriate for speech signals compared to the Gaussian model. The use of the Laplacian model has been proven to benefit the speech enhancement algorithms [48] [118]. The parameter estimation of the Laplacian model is crucial for the performance of speech enhancement. When the additive noise level is high, this limitation lowers the effectiveness of the enhancement algorithm.

1.2 Thesis Contribution

The contributions of this thesis are primarily the following:

- A recursive estimation for the Laplacian factor in existing algorithms in [48] [118] is developed. With this recursive approach, the estimation of the parameters can be improved leading to improved speech enhancement.

- An extension of the spectral subtraction is investigated by taking the power spectral to an extra power of α. We show that the appropriate value of α to achieve the best speech enhancement performance depends on the SNR. A generalized adaptive-α subtraction algorithm is proposed and is shown to work well for all the SNR circumstances.

- The loudness subtraction and over-subtraction approaches are proposed for speech enhancement applications. Instead of directly subtracting the whole noise loudness, the Laplacian statistical model of speech signals is considered to determine the proper parameters for this algorithm. The proposed algorithms outperform the corresponding spectral subtraction and over-subtraction when assessed by all tested quality measures.

- Finally, a loudness based maximum likelihood enhancement algorithm is presented. The proposed algorithm is compared with the MMSE algorithms and shown to provide comparable improvements with much lower complexity.

1.3 Thesis Organization

This thesis is organized as follows: Chapter 2 introduces the background knowledge related to this research. In Chapter 3, the proposed recursive speech enhancement employing Laplacian-Gaussian model is presented. In Chapter 4, the loudness subtraction type speech enhancement approach is generalized to an adaptive-α form. In Chapter 5, statistical model-based loudness subtraction and over-subtraction approaches are presented. In Chapter 6, a loudness-based speech enhancement algorithm using Maximum Likelihood (ML) criterion is investigated for both the Gaussian and Laplacian speech models.

Chapter 2

Background and Preliminaries of Speech Enhancement

In this chapter, we will review the background knowledge related to the work in this thesis.

2.1 Speech Decorrelation

Speech signals are highly correlated in the time domain. In speech processing applications, different transformations are applied to decompose the speech signal to uncorrelated domains for further processing [115].

In many applications, the speech signal is represented on a frame-by-frame basis. A frame of K speech samples in time domain is defined as:

$$\bar{x}(m) = [x(m), x(m-1), \ldots, x(m-K)]^T. \tag{2.1}$$

where $x(m)$ is the speech signal sample at time instant m. Also, let $\bar{y}(m)$ denote the

corresponding K-dimensional noisy speech signal. The noise vector is $\bar{n}(m)$, then

$$\bar{y}(m) = \bar{x}(m) + \bar{n}(m). \tag{2.2}$$

Assume $\bar{y}$ is a K-dimensional random vector with zero mean, then there exist a representation for $\bar{y}$ as follows,

$$\bar{y} = \sum_{i=1}^{r} Y_i w_i \qquad r \leq K \tag{2.3}$$

where $\{Y_i\}_{i=1}^{r}$ are zero mean and uncorrelated random variables, and $\{w_i\}_{i=1}^{r}$ are K-dimensional linearly independent basis vectors. The above equation can also be written as

$$\bar{y} = W\bar{Y}, \tag{2.4}$$

where $W \triangleq [w_1, w_2, \ldots, w_r]$ is a $K \times r$ matrix and $\bar{Y} \triangleq [Y_1, \ldots, Y_r]^T$ is a vector of r uncorrelated random variables. The set of all signal vectors X lie in a subspace of the Euclidean space $\mathbb{R}^K$ spanned by $\{w_i\}_{i=1}^{r}$. This subspace is referred to as the Signal Subspace.

Since the correlation between speech signals is commonly rather high, r is always chosen less than K, reflecting the fact that a speech data vector can be represented with a very small error by only a few principal components. Given W, these components can be computed. Further, the signal frame $\bar{y}$ can be reconstructed, approximately, as a linear combination of these few principal components.

In most transform-based speech enhancement approaches, the speech signals are first transformed into a less correlated domain, where enhancement is more feasible. Then the samples are modified by some classical enhancement process: spectral subtraction, Wiener filtering, etc.

In the following sections, we start by reviewing some of the main transformations used for speech enhancement.

2.1.1 Karhunen-Loève Transform (KLT)

KLT [1] [94] decomposes the signal into uncorrelated components by finding the eigen-decomposition of the short-time covariance matrix of the speech signals. Its transformation matrix depends on the input data. Let X denote a K-dimensional vector of signal samples, and the covariance matrix of X is denoted by

$$R_X \triangleq E\{XX^T\}. \tag{2.5}$$

Assume that the eigen-decomposition of R_X is as follows:

$$R_X \triangleq E\{XX^T\} = W\Lambda_X W^T, \tag{2.6}$$

where, $W \triangleq [w_1, w_2, \ldots, w_K]$ denotes an orthonormal matrix of eigenvectors of R_X, and $\Lambda_X \triangleq \text{diag}(\lambda_X(1), \lambda_X(2), \ldots, \lambda_X(K))$ is a diagonal matrix with its diagonal elements as the eigenvalues of R_X. Also assume that Λ_X are non-increasingly ordered as

$$\lambda_X(1) \geq \lambda_X(2) \geq \ldots \geq \lambda_X(K). \tag{2.7}$$

For a colored signal such as speech, the rank of R_X is less than K or some of the smaller eigenvalues are negligible. Thus, some of the eigenvalues in (2.7) can be assumed to be zero, which means the signal energy is negligible in some directions of the space. The transform W^H will optimally concentrate the signal power in a relatively small number of uncorrelated coefficients. In fact, we only need a small part of the eigen-pairs (λ_i, w_i) for $i = 1, \ldots, r$, with $r < K$ to represent the signal accurately and to reduce computation complexity and storage.

The sample vector X is represented by an r-dimensional vector S and a $K \times r$ transform matrix W. The reconstructed signal is represented by

$$\hat{X} = WS = WW^T X, \tag{2.8}$$

with $W = [w_1, w_2, \ldots, w_r]$ and $S = [s_1, s_2, \ldots, s_r]^T$.

The matrix W^T in (2.6) is called Karhunen-Loève transform (KLT). An important property of W^T is that the covariance matrix of $S = W^T X$ is diagonal. The column span of W corresponding to nonzero eigenvalues is the signal subspace.

In [40], the implementation of KLT is performed by eigenvalue decomposition (ED). The approach is not suitable for processing of non-stationary signals because it requires repeated ED, which is a very time consuming task. In order to overcome this difficulty, Yang developed a new type of KLT tracking algorithm called projection approximation subspace tracking [112]. This algorithm introduces an unconstrained cost function with a global minimum which corresponds to the desired signal subspace. This approach is fast, simple and has good performance in tracking process. Thus, it will be used in the following chapters.

2.1.2 Discrete Cosine transform (DCT)

DCT is an alternative for the optimal transform coding of Gaussian sources. It is a computationally effective harmonic transform that can whiten an autoregressive signal (speech) almost as well as KLT, if the data size is large enough [62].

For the given input vector $\bar{y}$, the DCT coefficients are calculated as:

$$Y_k(m) = \mathrm{w}_{\mathrm{DCT}}(k) \sum_{n=1}^{K} y(m-n+1) \cos \frac{k\pi(2n-1)}{2K}, \quad k = 1, \ldots, K \tag{2.9}$$

where

$$\mathrm{w}_{\mathrm{DCT}}(k) = \begin{cases} \sqrt{\frac{1}{K}}, & \text{for } k = 1 \\ \sqrt{\frac{2}{K}}, & \text{for } 2 \leq k \leq K. \end{cases} \tag{2.10}$$

The coefficients $Y_k(m)$ of the DCT are less correlated than the original speech $\bar{y}(n)$. They

compact the energy of a speech block into the DCT coefficients $\{Y_k\}_{k=1}^{K}$. The DCT coefficients can be used to recover the signal directly by

$$y(m-n+1) = \mathrm{w}_{\mathrm{DCT}}(n) \sum_{k=1}^{K} Y_k(m) \cos \frac{n\pi(2k-1)}{2K}, \quad n = 1, \ldots, K. \tag{2.11}$$

DCT has been shown to perform well in speech enhancement applications [103]. It is computationally efficient, has a signal independent transformation matrix and can whiten the inputs data almost as well as KLT. All these features lead to excellent performance of the algorithm.

2.1.3 Transformation for Speech Quality Assessment: Bark Frequency Transforms

The Bark frequency scale mimics the scale based on which the ear processes the received signal. The Bark scale has 24 Barks, corresponding to the first 24 critical bands of hearing.

Critical band width differs within the frequency range. The published Bark band edges [100] [101] are given in Hertz as [0, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500]. The published band centers in Hertz are [50, 150, 250, 350, 450, 570, 700, 840, 1000, 1170, 1370, 1600, 1850, 2150, 2500, 2900, 3400, 4000, 4800, 5800, 7000, 8500, 10500, 13500].

The bark frequency transform is widely used in the evaluation of the speech quality [15] [72], which will be shown later in this section. Also it has been used in speech coding [69] and speech enhancement [109].

2.2 Speech Enhancement Approaches

Over the past five decades, the problem of speech enhancement has been discussed by many researchers [80]. The main objective of enhancement is to improve the performance of speech communication systems in a noisy environment. Many applications, such as speech recognition, speech coding and hearing aids, will only have access to the noisy speech. Speech enhancement is usually necessary as the first step in these applications. Depending on the specific application, the objective of a single channel speech enhancement system may be to improve the quality, increase intelligibility, reduce listener fatigue, etc. Some systems may meet several of these objectives at the same time, while others may have to sacrifice one objective to meet the requirement of the other. Different algorithms of speech enhancement will be designed to meet specific objectives.

Most speech enhancement research focuses on removing the corrupting noise to improve the overall quality of the speech signal. In most algorithms, the noise term is assumed to be additive and uncorrelated to the clean speech signal. Some other types of practical noise can be transformed into additive noise [80]. For example, a multiplicative or convolutional noise can be converted to an additive noise by the homomorphic transformation [3].

Classical speech enhancement approaches include: spectral subtraction [80], MMSE-based algorithms [41], Wiener filtering [31], etc.. More recent approaches [38] incorporate the subspace processing [40], the auditory masking [109], adaptive filtering [113]. These approaches will be discussed in detail next.

2.2.1 Spectral Substraction

This approach estimates the Power Spectral Density (PSD) of the clean signal by subtracting the estimated PSD of the noise from the PSD of the noisy signal. Usually the PSD

is estimated with Short-Time Fourier Transform (STFT). Each estimate of PSD of noisy speech is performed within a short segment of signals because the short-time spectral amplitude is important for both speech quality and intelligibility. The PSD of the noise is typically obtained using several frames of noise-only segment [71] [80] [87] [99].

As the noise and the clean speech are assumed to be uncorrelated, an estimate of the clean speech PSD at the mth frame and frequency ω, $\hat{X}(m,\omega)$, is given by

$$\left|\hat{X}(m,\omega)\right|^2 = \begin{cases} |Y(m,\omega)|^2 - \hat{S}_N(\omega) & \text{if } |Y(m,\omega)|^2 - \hat{S}_N(\omega) \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

where $|Y(m,\omega)|^2$ is the measured noisy speech PSD and $\hat{S}_N(\omega)$ is the noise PSD estimate. The phase of the PSD is relatively unimportant in short-time processing. Thus, the measured phase of the noisy speech will be kept for the enhanced speech. Then the clean speech PSD estimate is given by:

$$\hat{X}(m,\omega) = \left|\hat{X}(m,\omega)\right| e^{j\angle Y(m,\omega)}. \tag{2.13}$$

This approach is simple in calculation and does not need any assumptions about the signal itself. The main drawback of this approach is the annoying musical noise remaining in the enhanced signal. This can be easily explained by the statistical nature of the noise. Noise estimation is based on an average of noise samples over time. When the estimate is higher than the actual noise sample in a frame, noise will be eliminated in that frame and sometimes the speech will be distorted. When the estimate is lower than the actual noise sample, a portion of the noise will still remain in the enhanced signal. The later case results in short sinusoids at random time and frequency sounding like a musical tone, hence the name the "musical noise".

2.2.2 Wiener Filtering

The Wiener filter [31] [44] [80] [86] is the optimal filter with respect to minimizing the mean squared error between the filter outputs and the desired clean signal. Given a number of observations, $y(m), y(m-1), \ldots, y(m-K)$, which are the sum of the desired signal $x(m), x(m-1), \ldots, x(m-K)$ and noise term, the estimate $\hat{x}(m)$ is obtained by a linear filtering applied on the set of observations:

$$\hat{x}(m) = \sum_{k=0}^{K} h_k y\,(m-k). \tag{2.14}$$

The Wiener filter is the optimum filter minimizing

$$E\left[e^2(m)\right] = E\left[(\hat{x}(m) - x(m))^2\right]. \tag{2.15}$$

The optimum filter is based on the principle of orthogonality that the error is orthogonal to the observations:

$$E\left[e(m)y(m)\right] = 0. \tag{2.16}$$

It can be shown that under the condition that $x(m)$ and $n(m)$ are uncorrelated and stationary, the coefficients of the filter that minimize the mean squared error must satisfy:

$$R_{x,y}(m) = \sum_{k=0}^{K} h_k R_y\,(m-k), \tag{2.17}$$

where $R_{x,y}$ and R_y are the cross-correlation and auto-correlation function, respectively. As we assume that the clean speech signal and noise are uncorrelated and zero-mean, then

$$R_{x,y}(m) = R_x(m) \tag{2.18}$$

and

$$R_y(m) = R_x(m) + R_n(m). \tag{2.19}$$

If we estimate the signal using the entire observation of y from time $-\infty$ to $+\infty$, the equation of the filter (2.17) becomes

$$R_{x,y}(m) = \sum_{k=-\infty}^{+\infty} h_k R_y(m-k). \tag{2.20}$$

The Fourier transform of above equation is

$$S_{x,y}(\omega) = H(\omega) S_y(\omega). \tag{2.21}$$

From the equations (2.18) and (2.19), we can derive:

$$S_{x,y}(\omega) = S_x(\omega) \tag{2.22}$$

and

$$S_y(\omega) = S_x(\omega) + S_n(\omega). \tag{2.23}$$

So the optimal Wiener Filter in frequency domain is

$$H(\omega) = \frac{S_x(\omega)}{S_x(\omega) + S_n(\omega)}. \tag{2.24}$$

Sometimes the term S_x can't be estimated directly from the noisy signal y, instead it is estimated with the noisy observation S_y and estimated noise S_n. Thus, equation (2.24) of the filter can be rewritten as:

$$H(\omega) = \frac{S_y(\omega) - S_n(\omega)}{S_y(\omega)}. \tag{2.25}$$

The parameterized Wiener filter is a extended version Wiener filter [20]:

$$H(\omega) = \left(\frac{S_x(\omega)}{S_x(\omega) + \alpha S_n(\omega)} \right)^{\beta}. \tag{2.26}$$

Different choices of the constants α and β lead to different filters. Usually α is called the noise suppression factor or the over-subtraction factor, and β is the power of the filter. The

power subtraction filter is obtained when $\beta = 0.5$ and $\alpha = 1$. The classical Wiener filter is when both β and α are 1.

The estimated noise power S_n is the average power of the noise samples and may be lower than the actual noise sample sometimes. Hence the noise suppression factor is usually chosen larger than 1 to better eliminate the residue noise at the cost of possible higher distortion.

One major disadvantage of Wiener filter is that it is designed to minimize the MSE. Given that several studies [54] have shown that the correlation between the MSE and speech quality is low, then minimizing MSE does not necessarily result in optimal speech quality.

2.2.3 Subspace Speech Processing

Subspace speech processing methods [56] [60] [65] [67] project the noisy speech signal onto the subspaces: signal-plus-noise subspace and noise only subspace. Then the processing of speech is done to the subspace signal components instead of the signal itself. The main research focus in this area is speech enhancement algorithms based on subspace methods. The decomposition of the signal into subspaces can be done with two methods: eigenvalue decomposition (EVD) [21] [68] [79] [94] or singular value decomposition (SVD) [22] [108].

2.2.3.1 Eigenvalue Decomposition (EVD) based SE

The EVD decomposes the covariance matrix $\mathbf{R}$ of the input signal vector $\mathbf{y}$, which is the combination of clean signal $\mathbf{x}$ and additive noise $\mathbf{n}$. Let $\mathbf{x}(m)$ be a K-dimensional vector of samples of clean speech $x(t)$ at time m as denoted by:

$$\mathbf{x}(m) = [x(m), x(m-1), \ldots, x(m-K+1)]^T . \tag{2.27}$$

Thus,

$$\mathbf{y}(m) = \mathbf{x}(m) + \mathbf{n}(m) \tag{2.28}$$

where $\mathbf{n}(m)$ is the K-dimensional noise vector at time m.

The covariance matrix of clean speech $\mathbf{x}(m)$ is defined as follows:

$$R_{\mathbf{x}}(m) = E[\mathbf{x}(m)\mathbf{x}^T(m)]. \tag{2.29}$$

Now, consider the eigen-decomposition of $R_{\mathbf{x}}(m)$ to be

$$R_{\mathbf{x}}(m) = W(m)\Lambda_{\mathbf{x}}(m)W^T(m) \tag{2.30}$$

where $\Lambda_X(m)$ is the diagonal matrix containing the eigenvalues of the clean speech covariance matrix $R_{\mathbf{x}}(m)$, and the unitary matrices $W^T(m)$ and $W(m)$ are called the KLT and the Inverse KLT (IKLT) of the clean signal $\mathbf{x}(m)$ [94]. Unlike the DCT, the matrix $W(m)$ depends on the signal to be decomposed.

For additive noise uncorrelated with the clean signal, the covariance matrix of $\mathbf{y}$ will be given by:

$$R_{\mathbf{y}}(m) = E[\mathbf{y}(m)\mathbf{y}^T(m)] = R_{\mathbf{x}}(m) + R_{\mathbf{n}}(m) = W(m)\left(\Lambda_{\mathbf{x}}(m) + \Lambda_{\mathbf{n}}(m)\right)W^T(m). \tag{2.31}$$

A subspace approach has been proposed [40] [59] to enhance noisy speech signals. The noisy signal is defined to have a form of

$$\mathbf{y} = \Psi\mathbf{s} + \mathbf{n} = \mathbf{x} + \mathbf{n} \tag{2.32}$$

where y, x, s and n represent the noisy speech, clean speech, the uncorrelated representation of clean speech and noise only signals, respectively. The set of all possible signal vectors $\{\mathbf{x}\}$ will lie in a subspace of the Euclidean space spanned by the columns of Ψ. This

subspace is referred as "signal subspace". Now the speech enhancement problem becomes: find a matrix H to form a linear estimator $\hat{\mathbf{x}} = H\mathbf{y}$ under some constraint.

The error of the above estimator is:

$$\varepsilon = \hat{\mathbf{x}} - \mathbf{x} = (H - I)\,\mathbf{x} + H\mathbf{n} = \varepsilon_{\mathbf{x}} + \varepsilon_{\mathbf{n}} \tag{2.33}$$

where the first term represents the speech distortion and the second term represents the residual noise. Then the time-domain constrained optimization problem becomes:

$$\min_{H} \hat{\varepsilon_{\mathbf{x}}^2} \tag{2.34}$$

subject to:

$$\frac{1}{K}\hat{\varepsilon_{\mathbf{n}}^2} \leq \sigma_{\mathbf{n}}^2. \tag{2.35}$$

For the white noise case, $R_{\mathbf{n}} = \sigma_{\mathbf{n}}^2 I$, the solution is given by [40]:

$$H_{opt} = R_{\mathbf{x}} \left(R_{\mathbf{x}} + \mu R_{\mathbf{n}}\right)^{-1} \tag{2.36}$$

where $R_{\mathbf{x}}$ and $R_{\mathbf{n}}$ are the covariance matrices of the clean speech and noise signals, respectively, and μ is the Lagrange multiplier. It has been shown that varying μ can control the tradeoff between the residual noise and speech distortion [59].

$R_{\mathbf{x}} = U\Delta_{\mathbf{x}}U^T$ is the eigen-decomposition of $R_{\mathbf{x}}$. U is the unitary eigenvector matrix and $\Delta_{\mathbf{x}}$ is the diagonal eigenvalue matrix of $R_{\mathbf{x}}$. The solution can be rewritten as

$$H_{opt} = U\Delta_{\mathbf{x}} \left(\Delta_{\mathbf{x}} + \mu {\sigma_{\mathbf{n}}}^2 I\right)^{-1} U^T \tag{2.37}$$

In the algorithm by Rezayee [94], the matrix $U^T R_{\mathbf{n}} U$ was approximated by a diagonal matrix $\Delta_{\mathbf{n}}$ with different diagonal components to allow for the case of colored noise. Then, the solution is modified to

$$H_{opt} = U\Delta_{\mathbf{x}} \left(\Delta_{\mathbf{x}} + \mu\Delta_{\mathbf{n}}\right)^{-1} U^T. \tag{2.38}$$

It should be noted that the matrix $U^T R_{\mathbf{n}} U$ is not exactly a diagonal matrix because U is designed to diagonalize $R_{\mathbf{x}}$ and not $R_{\mathbf{n}}$. In [59], a matrix is constructed to diagonalize $R_{\mathbf{x}}$ and $R_{\mathbf{n}}$ simultaneously:

$$V^T R_{\mathbf{x}} V = \Lambda_{\mathbf{x}} \tag{2.39}$$

$$V^T R_{\mathbf{n}} V = I \tag{2.40}$$

where $\Lambda_{\mathbf{x}}$ and V are the eigenvalue and eigenvector matrix of $\Sigma = R_{\mathbf{n}}^{-1} R_{\mathbf{x}}$, respectively. It can be shown that $\Lambda_{\mathbf{x}}$ is a real matrix. Note that unlike U in the previous assumption, the eigenvector matrix V is not orthogonal. Applying it to (2.36), the modified optimal linear estimator is

$$\begin{aligned} H_{opt} &= R_{\mathbf{n}} V \Lambda_{\mathbf{x}} \left(\Lambda_{\mathbf{x}} + \mu I\right)^{-1} V^T \\ &= V^{-T} \Lambda_{\mathbf{x}} \left(\Lambda_{\mathbf{x}} + \mu I\right)^{-1} V^T \end{aligned}$$

This approach is a generalization of the Ephraim's approach in [40] without restricting the noise to be white.

Hu [59] proposed a variable μ to improve the trade-off between the speech distortion and residual noise, which is a focus topic in many speech enhancement papers.

2.2.3.2 Singular Decomposition (ESVD) based SE

The SVD is used to decompose the signal matrix with the following form:

$$\tilde{X}(m) = \begin{pmatrix} x(m) & x(m+1) & \dots & x(m+K-1) \\ x(m+1) & x(m+2) & \dots & x(m+K) \\ \vdots & \vdots & & \vdots \\ x(m+L-1) & x(m+L) & \dots & x(m+K+L-2) \end{pmatrix} \tag{2.41}$$

when $K + L - 1$ samples of the signal are available ($L \geq K$) [56]. As before, $\tilde{X}$ is used to represent the clean signal and the noisy signal is

$$\tilde{Y}(m) = \tilde{X}(m) + \tilde{N}(m). \tag{2.42}$$

Moreover, $\tilde{X}$ is rank deficient with rank $J < K$ while $\tilde{Y}$ and $\tilde{N}$ have full rank K. Since the noise is assumed to be broadband and zero-mean.

The SVD is defined by the following theorem.

Theorem If $\tilde{Y} \in \mathbb{R}^{L \times K}$ with $L \geq K$, then there exist matrices

$$U = [u_1 \dots u_K] \in \mathbb{R}^{L \times K} \tag{2.43}$$

and

$$V = [v_1 \dots v_K] \in \mathbb{R}^{K \times K} \tag{2.44}$$

with orthonormal columns such that

$$\tilde{Y} = U \Sigma V^T = \sum_{i=1}^{K} u_i \sigma_i v_i^T \tag{2.45}$$

where $\Sigma = \text{diag}\,(\sigma_1, \sigma_2, \dots, \sigma_K)$ with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_K \geq 0$.

The diagonal elements of Σ are called the singular values of $\tilde{Y}$ and their set is called the singular spectrum. The columns of U and V are called the left and right singular vectors.

For the rank deficient matrix $\tilde{X}$ with rank J, it is possible to partition the SVD of $\tilde{X}$ as

$$\tilde{X} = U_{\tilde{X}} \Sigma_{\tilde{X}} V_{\tilde{X}}^T = [U_{\tilde{X}1}\, U_{\tilde{X}2}] \begin{bmatrix} \Sigma_{\tilde{X}1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_{\tilde{X}1}^T \\ V_{\tilde{X}2}^T \end{bmatrix} \tag{2.46}$$

where $U_{\tilde{X}1} \in \mathbb{R}^{L \times J}$, $V_{\tilde{X}1} \in \mathbb{R}^{K \times J}$ and $\Sigma_{\tilde{X}1} = \text{diag}\,(\sigma_{\tilde{X}1}, \sigma_{\tilde{X}2}, \dots, \sigma_{\tilde{X}J})$.

Now we use the above SVD for the application of speech enhancement. The SVD of noisy signal $\tilde{Y}$ is also partitioned as following:

$$\tilde{Y} = U_{\tilde{Y}}\Sigma_{\tilde{Y}}V_{\tilde{Y}}^T = [U_{\tilde{Y}1}\ U_{\tilde{Y}2}] \begin{bmatrix} \Sigma_{\tilde{Y}1} & 0 \\ 0 & \Sigma_{\tilde{Y}2} \end{bmatrix} \begin{bmatrix} V_{\tilde{Y}1}^T \\ V_{\tilde{Y}2}^T \end{bmatrix} \tag{2.47}$$

and the SVD of the clean speech signal $\tilde{X}$ is given by (2.46).

First the following assumptions are made:

1. The signal is orthogonal to the noise in the sense $E\{\tilde{X}^T\tilde{N}\} = 0$.
2. The noise is white, i.e., $E\{\tilde{N}^T\tilde{N}\} = \sigma_{\tilde{N}}^2 I$
3. The smallest singular value of $\Sigma_{\tilde{Y}1}$ is strictly larger than the largest singular value of $\Sigma_{\tilde{Y}2}$ in the SVD of $\tilde{Y}$.

With these assumption, the SVD of $\tilde{Y}$ can be written in terms of SVD of $\tilde{X}$:

$$\begin{aligned} \tilde{Y} &= \tilde{X} + \tilde{N} \\ &= U_{\tilde{X}1}\Sigma_{\tilde{X}1}V_{\tilde{X}1}^T + \tilde{N}V_{\tilde{X}1}V_{\tilde{X}1}^T + \tilde{N}V_{\tilde{X}2}V_{\tilde{X}2}^T \\ &= \left[\left(U_{\tilde{X}1}\Sigma_{\tilde{X}1} + \tilde{N}V_{\tilde{X}1}\right)\left(\Sigma_{\tilde{X}1}^2 + \sigma_{\tilde{N}}^2 I\right)^{-1/2} \quad \left(\sigma_{\tilde{N}}^{-1}\tilde{N}V_{\tilde{X}2}\right)\right] \\ &\quad \begin{bmatrix} \left(\Sigma_{\tilde{X}1}^2 + \sigma_{\tilde{N}}^2 I\right)^{1/2} & 0 \\ 0 & \sigma_{\tilde{N}} I \end{bmatrix} \begin{bmatrix} V_{\tilde{X}1}^T \\ V_{\tilde{X}2}^T \end{bmatrix} \end{aligned} \tag{2.48}$$

$$\tilde{Y} = [U_{\tilde{Y}1}\ U_{\tilde{Y}2}] \begin{bmatrix} \Sigma_{\tilde{Y}1} & 0 \\ 0 & \Sigma_{\tilde{Y}2} \end{bmatrix} \begin{bmatrix} V_{\tilde{Y}1}^T \\ V_{\tilde{Y}2}^T \end{bmatrix} \tag{2.49}$$

With the above assumptions and $U_{\tilde{X}}$, $V_{\tilde{X}}$ have orthonormal columns,

$\left[\left(U_{\tilde{X}1}\Sigma_{\tilde{X}1} + \tilde{N}V_{\tilde{X}1}\right)\left(\Sigma_{\tilde{X}1}^2 + \sigma_{\tilde{N}}^2 I\right)^{-1/2} \quad \left(\sigma_{\tilde{N}}^{-1}\tilde{N}V_{\tilde{X}2}\right)\right]$ have orthonormal columns. Equation (2.49) is the SVD of $\tilde{Y}$. It has the following relationships with the SVD of X:

$$\Sigma_{\tilde{Y}1} = \left(\Sigma_{\tilde{X}1}^2 + \sigma_{\tilde{N}}^2 I\right)^{1/2} \tag{2.50}$$

$$\Sigma_{\tilde{Y}2} = \sigma_{\tilde{N}} I$$

$$U_{\tilde{Y}1} = \left(U_{\tilde{X}1}\Sigma_{\tilde{X}1} + \tilde{N}V_{\tilde{X}1}\right)\left(\Sigma_{\tilde{X}1}^2 + \sigma_{\tilde{N}}^2 I\right)^{-1/2} \tag{2.51}$$

$$= \left(U_{\tilde{X}1}\Sigma_{\tilde{X}1} + \tilde{N}V_{\tilde{X}1}\right)\Sigma_{\tilde{Y}1}^{-1} \tag{2.52}$$

$$U_{\tilde{Y}2} = \left(\sigma_{\tilde{N}}^{-1}\tilde{N}V_{\tilde{X}2}\right)$$

$$V_{\tilde{Y}1}^T = V_{\tilde{X}1}^T \tag{2.53}$$

$$V_{\tilde{Y}2}^T = V_{\tilde{X}2}^T \tag{2.54}$$

$$\tag{2.55}$$

Under the above assumptions, the row space of $\tilde{X}$ (represented by $V_{\tilde{X}1}^T$ and $V_{\tilde{X}2}^T$) can be estimated consistently, which means the estimation will converge in probability as $L \rightarrow \infty$. Since the column space of $\tilde{X}$ (represented by $U_{\tilde{X}1}$) can't be recovered from the the SVD of $\tilde{Y}$, not even asymptotically, the least-square [34] [66] and minimum variance [37] estimates are used instead to estimate $\tilde{X}$ from $\tilde{Y}$.

The noise reduction algorithm based on the Singular Value Decomposition (SVD) is a robust and widely used computational tool in noise suppression techniques. The problem is that this method deals only with white noise and the LS estimate is sensitive to the number of retained singular values. In [66], a noise reduction method based on the Quotient Singular Value Decomposition (QSVD) is presented with pre-whitening as one part of the algorithm. Moreover, by using a Minimum Variance (MV) estimate [37] of the signal-only matrix, the algorithm is less sensitive to the choice of retained singular values.

In [40], the proposed estimators attempt to improve the quality of the noisy signal while

minimizing any loss in its intelligibility. The focus is on the Linear Minimum Mean Square Error (LMMSE) estimation criterion, which minimizes the error between the enhanced and the clean signal. The optimal estimator in this sense is the well-known Wiener filter. However, the error signal represents both signal distortion and residual noise, which cannot be simultaneously minimized. Thus, the proposed estimators [32] control the level of the perceptually harmful residual noise (musical noise), while minimizing the signal distortion.

To conclude this section, we note that subspace methods decorrelate the time-varying speech signals and allow the speech enhancement algorithm to work with un-correlated speech components in the transformed space. One major drawback, though, is that tracking of the subspace is usually done in noisy circumstances. As the speech can only be assumed to be statistically stable in 10-40 ms, the accurate decomposition is very critical.

2.2.4 Use of Human Auditory Systems in Speech Enhancement

The masking property of the human hearing system can be utilized to improve the performance of different speech processing applications.

It has been used with spectral subtraction [74] and subspace methods [73]. Virag [109] proposed a single channel speech enhancement system based on masking properties of the human auditory system. This system was based on the generalized spectral subtraction. It calculated a noise masking threshold, below which all additive noise components were inaudible. Then the noise in each band was lowered to the level that just below the masking threshold instead of eliminating it all. It overcame the limitations of one channel subtraction-type enhancement system in additive background noise at low SNR's. This speech enhancement algorithm was superior to the classical methods, especially at the low SNR's, since it introduces less distortion by applying the masking threshold.

2.3 The Loudness Model

The loudness model was originally proposed by Zwicker [10] [19]. It is used to simulate the signal as perceived by the human auditory systems. The speech signal is pre-filtered to simulate the outer and middle ear transmission and then separated into critical bands. Within each band, the signal is transformed to the representation in the loudness domain with the knowledge of the excitation pattern. The loudness of the signal is the summation of the loudness across the frequency bands.

The bands are defined based on the Bark scale rather than a linear frequency scale due to the acoustic properties of the human ear. The excitation pattern will be calculated from a series of auditory filters. The shapes of the auditory filters is decided using the empirical formula

$$W(g) = (1 + gl)\exp(-gl) \tag{2.56}$$

where g is the normalized frequency deviation from the center frequency of the filter $g = \frac{|f-f_0|}{f_0}$ and l determines the bandwidth and the slopes of the skirts of the filters. This filter shape has been called the "rounded exponential filter" or the ROEX filter [88].

First, we define the intensity of the speech signal [10]. Intensity is a measure of the time averaged energy flux of a speech signal. It is a vector that has the units of power divided by area. When the energy of the speech is radiated uniformly and no loss, the intensity at the point with distance r from the source can be represented in the following equation:

$$|I| = \frac{E}{4\pi r^2}. \tag{2.57}$$

where E is the energy of sound at the source and r is the distance from the source to receiver.

The intensity can also be defined by the sound pressure p, which is directly caused

by the sound wave. The sound pressure has values from 10^{-5} Pa (absolute threshold of hearing) and 10^{2} Pa (threshold of pain) in psychoacoustics [10]. Here the Pa (Pascal) is the same unit used for atmospheric pressure.

Usually the sound pressure level P [10] is used. It is defined as

$$P = 20 \log (p/p_0)\, dB. \tag{2.58}$$

The reference value of the sound pressure p_0 is selected as $p_0 = 20\mu$Pa. Also, P can be represented by the intensity I:

$$P = 10 \log (I/I_0)\, dB. \tag{2.59}$$

The reference I_0 is $10^{-12} W/m^2$.

The loudness model is applied to calculate a specific loudness N' in a band from the excitation in that band. An absolute intensity threshold I_{ThQ} is defined for internal noise that masks the speech at very low levels. The intensity below the threshold is inaudible.

There are several different laws for loudness models. The simplest one is Steven's Law. It assumes that human intensity sensation grows with physical intensity according to a power law. There are several different laws with these assumptions.

1. Steven's power law (Stevens [104], 1957):

$$N' = C \cdot I^{\alpha} \quad . \tag{2.60}$$

2. Uncompressed internal noise power law (Stevens [105], 1966):

$$N' = C \cdot (I - I_{ThQ})^{\alpha} . \tag{2.61}$$

3. Compressed internal noise power law (Humes and Jesteadt [63], 1991):

$$N' = C \cdot \left(I^{\alpha} - I_{ThQ}^{\alpha}\right) . \tag{2.62}$$

4. Zwicker's mixed internal noise power law (Zwicker [10], 1965):

$$N' = C \cdot ((I + k_1 I_{ThQ})^{\alpha} - (k_2 I_{ThQ})^{\alpha}) . \tag{2.63}$$

In the above equations, C, k_1 and k_2 are constants and α is a compressive exponent assumed to be in the range of 0.23-0.3 for normal-hearing subjects. For example, it is chosen to be 0.27 in [58] [93] and 0.3 in [23]. Clearly, the loudness N' is a non-linear, compressive function of the stimulus energy I.

The overall loudness of speech is the integration of the specific loudness N' across the frequency bands:

$$N = \int N'(z)dz \tag{2.64}$$

where z is the frequency on a Bark or ERB (Equivalent Rectangular Bandwidth) scale [100].

2.4 Speech Quality Assessments

Following the enhancement of a speech signal, we need to evaluate the quality of the enhanced speech. Speech quality can be evaluated subjectively by using the listening test by human subjects. Since degradations are introduced by the environment and the processing algorithms, it is difficult for any one listener to measure the quality of speech signals over a wide range of different kinds of distortions. It is also very costly to conduct such tests.

For these practical reasons, we need an objective measurement technique for the subjective speech quality [6] [110]. Typically the enhanced signal is compared to the original reference signal based on some measure that indicates the quality of the enhanced signal. This objective measure can provide an immediate and reliable estimate of the quality of the speech enhancement algorithm.

The block diagram of a typical speech quality assessment method is shown in Figure 2.1. Usually, a preprocessing step is applied prior to the quality measure. The preprocessing may include serial to parallel conversion, transformation and etc..

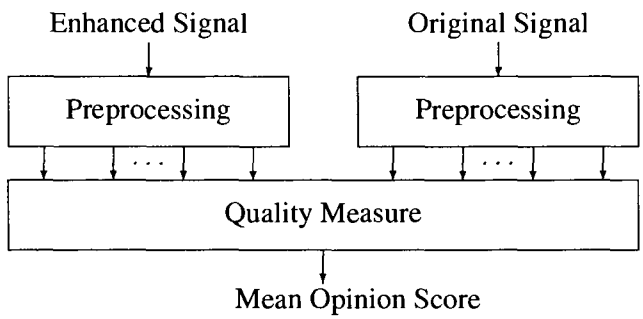

Figure 2.1: The block diagram of a typical speech quality assessment system

Mean Opinion Score (MOS) is the mean of the opinion scores of a speech signal given by many listeners measuring the speech subjective quality. Objective measures described above must be as correlated as possible to the MOS to be useful. The objective quality measures can be classified into two types: intrusive or non-intrusive.

2.4.1 Intrusive Speech Quality Measures

Intrusive speech quality measures assume that the clean reference signal is available. We will present several classes of these intrusive measures [24] [35] [78] based on the specific function used to assess quality.

SNR type measures

One relatively simple way to assess the quality of a signal is to estimate its SNR. This

is applied to the time domain signal samples as:

$$SNR = \frac{\sum_{n=1}^{N} x^2(n)}{\sum_{n=1}^{N} [x(n) - y(n)]^2} \tag{2.65}$$

where $x(n)$ is the clean speech signal and $y(n)$ is the corrupted signal after enhancement.

The segmental SNR is an alternative frame-by-frame measure of SNR and is defined as [24]:

$$SNR_{seg} = \frac{1}{M} \sum_{m=1}^{M} \log \left\{ 1 + \frac{\sum_{n=1}^{N} x^2(n)}{\sum_{n=1}^{N} [x(n) - y(n)]^2} \right\} \tag{2.66}$$

where M is the number of frames and N is the frame length.

Spectral distance type measures

In this case, the difference between $y(n)$ and $x(n)$ is measured in a spectral-type domain. One popular measure is the perceptual linear prediction (PLP) cepstral distance [24] used in the PLP domain and defined as:

$$CD = \sum_{n=1}^{P} [c_x(n) - c_y(n)]^2 \tag{2.67}$$

where $c_x(n)$ and $c_y(n)$ are the cepstral values of the signals $x(n)$ and $y(n)$ and P is the number of the cepstral coefficients.

Another such measure is the weighted slope spectral (WSS) distance [24] defined as:

$$WSS = \sum_{k=1}^{K} a * [(S_y(k+1) - S_y(k)) - (S_x(k+1) - S_x(k))]^2 \tag{2.68}$$

where S_x and S_y is the power spectra of x ,y, respectively, and a is a weighting function, which weights spectral peaks higher than the spectral valleys. K is the number of critical

bands that is used to calculate the power spectra. In [76], 36 critical bands are used.

Log spectral distance type measures

It is well known that the human ear is logarithmic in its response to sounds, i.e., the perceived loudness of a sound is proportional to the logarithm of the sound's intensity. Thus, the log spectral distance measure has been used to assess the relative quality of the signal.

The log spectral distance [24] is defined as:

$$SD = 10\log\left\{\frac{1}{K}\sum_{k=1}^{K}[S_y(k) - S_x(k)]^2\right\} \tag{2.69}$$

where $S_y(k)$ and $S_x(k)$ are the power spectra of the corrupted and clean speech signal in frequency ban k, respectively. k is the number of the critical bands considered in this measure. For example, this log spectral distance can also be defined in the Bark frequency scale spectrum where $k = 24$.

Itakura Distance

This distortion measure is defined based on the linear prediction of speech. The linear predictor estimates the future values of a signal x based on the linear combination of the previous samples:

$$\hat{x}_{\mathrm{n}} = \sum_{i=1}^{p} a_i x_{n-i}. \tag{2.70}$$

where a_i is the i^{th} LPC coefficient for $x(n)$. The residual total squared error is:

$$\sum_{n=0}^{N-1} e_n^2 = \sum_{n=0}^{N-1} \left(x_n - \sum_{i=1}^{p} a_i x_{n-i} \right)^2 \tag{2.71}$$

$$= \sum_{n=0}^{N-1} \left(x_n^2 - 2\sum_{i=1}^{p} a_i x_{n-i} x_n + \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j x_{n-i} x_{n-j} \right) \tag{2.72}$$

$$= \sum_{n=0}^{N-1} x_n^2 - 2\sum_{i=1}^{p} a_i \left(\sum_{n=0}^{N-1} x_n x_{n-i} \right) + \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j \left(\sum_{n=0}^{N-1} x_{n-i} x_{n-j} \right) \tag{2.73}$$

$$= R_{00} - 2\sum_{i=1}^{p} a_i R_{0i} + \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j R_{ij} \tag{2.74}$$

$$= \begin{bmatrix} -1 & a_1 & a_2 \ldots a_p \end{bmatrix} \begin{bmatrix} R_{00} & R_{01} & R_{02} & \ldots & R_{0p} \\ R_{10} & R_{11} & R_{12} & \ldots & R_{1p} \\ R_{20} & R_{21} & R_{22} & \ldots & R_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{p0} & R_{p1} & R_{p2} & \ldots & R_{pp} \end{bmatrix} \begin{bmatrix} -1 \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \tag{2.75}$$

The Itakura distance [64] is defined as:

$$IS = \frac{a_y^T R_x a_y}{a_x^T R_x a_x} \tag{2.76}$$

where a_y and a_x are the $p \times 1$ vector of LPC coefficients for the corrupted signal y and clean speech signal x, respectively. R_x is the autocorrelation matrix of the clean speech signal. $a_x^T R_x a_x$ is the minimum energy of the linear prediction error when the matrix R_x is available. The term $a_y^T R_x a_y$ is the energy of the prediction error when the corrupted signal y is used in place of x to calculate the LPC coefficients calculation.

The Itakura distance can also be represented in the frequency domain:

$$IS = \int_{-\pi}^{\pi} \left| \frac{Y\left(e^{j\omega}\right)}{X\left(e^{j\omega}\right)} \right|^2 \frac{d\omega}{2\pi} \tag{2.77}$$

which measures the difference between the spectral magnitudes of the corrupted and clean speech signals.

Log Likelihood Ratio (LLR)

The Log Likelihood Ratio (LLR) [6] is defined as:

$$LLR = \log\left(\frac{a_y^T R_x a_y}{a_x^T R_x a_x}\right) \tag{2.78}$$

which is an alternative of the previous Itakura Distance measure.

All the above traditional objective measures were compared with the subjective MOS measure in [24] [28]. The correlation between the objective and subjective measures was calculated. Since the quality of speech must be evaluated for intelligibility, background noise level and etc., it was concluded that no single measure can be used for all different cases. It has been suggested to combine some of these measures into a composite objective measure to better assess the quality.

Other intrusive measures, based on these classical measures, are given in [95] [110].

2.4.1.1 Perceptual Evaluation of Speech Quality (PESQ)

The first standardized measure is found in a competition by ITU-T and adopted as the ITU-T P.861. It is known as Perceptual Speech Quality Measure (PSQM). It measures the noise disturbance density, which is the absolute difference between the perceived loudness and reference loudness. It performs poorly with the telephone network, where packet loss and variable delay are very common.

In [110], the subjective loudness is used to evaluate the objective quality. The clean (reference) and the corrupted signals are pre-processed with critical-band filtering and the spectral energies at each band are perceptually weighted. This weighting process equalizes

the loudness at different frequencies. Then the loudness is calculated within each band. The mean square error (MSE) of the loudness between the clean and corrupted signals is calculated within each band and summarized for all the critical bands. This MSE is then normalized by the average loudness of the original speech to form the Bark spectral distortion (BSD) measure. This BSD model is later developed to the Perceptual Analysis Measurement System (PAMS) by Rix and Hollier [96].

From 1998 to 2000, the extended version of PSQM and PAMS were integrated and standardized. The combined model is the ITU-T recommendation P.862–the Perceptual Evaluation of Speech Quality (PESQ) [15] [97]. It is designed to simulate the human auditory system and provide a measure of the speech quality that is closer to the MOS score by subjective listening test. This recommendation provides an objective method for evaluating the subjective quality of narrow-band handset telephony and narrow band speech codec.

The PESQ compares a reference signal $x(m)$ with a degraded signal $y(m)$, which is the result of $x(m)$ passing through a communication system. The PESQ score is based on a number of tests that provide a combination of factors such as filtering, variable delay, coding distortions and channel errors. It is an algorithm to calculate the average distortion over time. The idea of the process is to transform both $x(m)$ and $y(m)$ to an internal representation that is similar to the psychoacoustic representation in the human auditory system.

The reference and degraded signal are sampled at either 8 kHz or 16 kHz sampling rate. These samples are transformed to the frequency domain using FFT with a Hanning window of 32 ms with 50% overlap. Both the reference and degraded signals are aligned in time and both signals are brought to the same power level. Auditory transformations are applied

to derive the psychoacoustic model of both signals.

Then the outputs are analyzed to determine the absolute difference between the degraded and the reference signals in the perceptual domain.

The power densities of $E_{X,WIRSS,n}$, which represented the power density of reference signal x after windowing and intermediate reference system (IRS) filtering, is transformed to the loudness domain using Zwicker's law [10]:

$$LX(f)_n = C \cdot \left(\frac{E_{ThQ,n}}{0.5}\right)^{\alpha} \left[\left(0.5 + 0.5\frac{E_{X,WIRSS,n}}{E_{ThQ,n}}\right)^{\alpha} - 1\right], \tag{2.79}$$

where $E_{ThQ,n}$ is a pre-defined absolute threshold.

And the disturbance, which is noted as $D(f)_n$, is derived as:

$$D(f)_n = LX(f)_n - LY(f)_n, \tag{2.80}$$

where $LY(f)_n$ is the loudness representation of the degraded signal in the loudness domain. Since humans can only hear differences of sounds above a certain threshold due to the masking property of the ear, only differences above this threshold are considered.

The PESQ integrates the disturbance first over the frequency scale and then averages over time. There are two parameters used: the symmetric disturbance (d_{SYM}) and the asymmetric disturbance (d_{ASYM}).

$$d_{SYM,n} = \sqrt[3]{\sum_f \left(|D(f)_n| \, W_f\right)^3} \tag{2.81}$$

and

$$d_{ASYM,n} = \sum_f \left(|DA(f)_n| \, W_f\right) \tag{2.82}$$

where f is summed over the Bark frequency bands, W_f is a constant proportional to the width of the Bark frequency bins. $DA(f)_n$ is the asymmetric disturbance density that is

derived by multiply $D(f)_n$ with an asymmetric factor. This asymmetry factor equals the ratio of the degraded and reference signal power densities raised to the power of 1.2. If the asymmetry factor is less than 3 it is set to zero. If it exceeds 12 it is clipped at that value. This factor emphasizes the disturbance of large additive noise compared to the reference signal in a time-frequency cell.

These two parameters, d_{SYM} and d_{ASYM}, can provide good accuracy of objective quality prediction. The MOS prediction in the PESQ standard uses the following mapping which is derived by training with sample database [15] [95]:

$$PESQMOS = 4.5 - 0.1d_{SYM} - 0.0309d_{ASYM}. \tag{2.83}$$

In other words, the PESQ score is mapped to a MOS-like scale, which is a single number within the range of -0.5 to 4.5.

As shown in [95], the PESQ score is compared to the subjective MOS in the tests of correlation coefficients and residual error distribution. The PESQ method clearly outperforms the traditional objective measures and the previously used objective speech quality assessment models such as the perceptual speech quality measure (PSQM) and the measuring normalized blocks (MNB). The PSQM and the MNB were only recommended for use in pure narrow-band codec assessment, while the PESQ can be used for wider range of real conditions including filtering, variable delays, coding distortion and channel errors.

2.4.2 Non-intrusive Speech Quality Measures

The above methods are intrusive measures in the sense that they require access to the clean speech signal to measure the relative quality of the corrupted signal. In some circumstance, the clean speech signal is not available. The non-intrusive measures have been proposed for these cases.

The Auditory Non-Intrusive QUality Estimation (ANIQUE) model is proposed in [75]. The speech signal is pre-filtered with cochlear filter-banks. The temporal envelope in each band is estimated. The change of the envelope is directly related to the quality of the speech. The ANIQUE then takes the envelope and transforms it with Hamming windowing and FFT into the "modulation spectrum". The ratio of the modulation spectrum to the noise spectrum is called the articulation-to-nonarticulation ratio. The ratios at each frequency and at each frame are accumulated to yield a measure of the objective quality of the speech.

A data mining approach is presented in [116]. The first step is the feature extraction. The signal is decomposed to 7 sub-bands and a distortion measure is calculated over time and bands. The L_2 norm of each measure is called a feature and a total of 209 features are available for data mining. The training of the model uses some clean-degraded speech pairs and finds the best subset of the features for the quality estimator. Then the speech can be evaluated with the estimator using a small subset of the features without need to access the clean speech.

Another non-intrusive model is proposed in [43]. A Gaussian mixture model is used to model the PLP (Perceptual Linear Prediction) coefficients of the speech signal. The Expectation-maximization algorithm is used to train the model. Then the consistency between the observation and the model is calculated. The consistency measure is mapped to objective MOS using multivariate adaptive regression splines.

In May of 2004, a non-intrusive objective speech quality measurement algorithm, Single-Ended Assessment Model (SEAM), was released as ITU-T standard P.563 [11] [97]. This standard is the first non-intrusive measurement that estimates the full range of distortions to predict the speech quality on a perception-based scale.

The P.563 begins with the model of the telephone handset and a Voice Activity Detector

(VAD) right after. This VAD identifies the speech portion and calculates the speech level. The speech is then investigated for three main classes of distortion: vocal tract analysis and unnaturalness of speech, analysis of strong additional noise and interruptions, mutes and time clipping.

The vocal tract analysis separates the voice from the non-speech intervals. Then it uses higher order statistical analysis to measure how human-like this speech is. The speech signal is classified to male and female voices in this section. It is then marked with active and non-active intervals by VAD. The active intervals are used to estimate the statistics of the speech based on the higher order statistical evaluation of the cepstral and LPC analysis.

The vocal tract analysis also estimates the periodicity in the signal and rates it, since the high periodic tone such as the DTMF signal is a very annoying disturbance. A more detailed description of the speech is given by comparing the speech with a pseudo reference signal generated by a speech enhancer. This comparison is done with P.862 like measurement.

The analysis of the additional noise is very critical. This section decides if the additional noise is the main degradation. If yes, the type of the noise is classified as static (noise power not correlated with the speech) or dependent (noise power depends on signal power envelope. Then the noise parameters are estimated based on different characteristics of the noise.

The P.563 algorithm can detect the interruptions and mutes during the speech. It can distinguish the word ends and abnormal signal interruptions as well as unnatural silence intervals. All of these are used in the evaluation of the quality.

Overall, the P.563 algorithm decides a distortion class for the speech intervals. At each interval, it evaluate the objective quality according to the distortion class. Then it gives the

overall calculation of the objective speech quality.

2.4.3 Objective Measures for Speech Enhancement in this Thesis

It has been noticed that PESQ itself is not enough for the evaluation of speech enhancement algorithms. It is also suggested additional measures for assessing noise removal and distortion are needed [61]. In this thesis, all three aspects of the enhanced speech will be evaluated for comparison. The PESQ is used since it is the most correlated measure to the MOS scores. Segmental SNR will be used to measure the residue noise in the enhanced signals. Log likelihood ratio will be used to measure the distortion.

In 2003, ITU-T standardized the methodology for evaluating the noise suppression algorithms [14]. The main reason to develop such a standard is that the speech enhancement algorithms usually compromise between two aspects: the noise suppression and the distortion to the speech. Under such circumstance, the listener can be confused of which one should be their focus when evaluate a speech enhancement algorithm. So the ITU-T recommendation P.835 require the listener to evaluate the speech signal, the background noise and the overall quality separately.

In [61], evaluation of objective measures for speech enhancement is discussed. The three quality measures required by ITU-T P.835 can be estimated by the measures presented above. The linear approximation can be given by:

$$M_{sig} = 3.093 - 1.029 \cdot LLR + 0.603 \cdot PESQ - 0.009 \cdot WSS \tag{2.84}$$

$$M_{back} = 1.634 + 0.063 \cdot segSNR + 0.478 \cdot PESQ - 0.007 \cdot WSS \tag{2.85}$$

$$M_{overall} = 1.594 - 0.512 \cdot LLR + 0.805 \cdot PESQ - 0.007 \cdot WSS \tag{2.86}$$

where M_{sig}, M_{back} and $M_{overall}$ are the evaluations of the speech signal, the background

noise and the overall quality, respectively. Overall, the objective evaluation of speech enhancement does not have a standardized measure yet. Further research is currently underway.

2.5 Chapter Summary

In this Chapter, some basic aspects of speech transformation, speech enhancement, the loudness model of speech and speech quality assessment were reviewed. As the PESQ is based on the loudness disturbance measure of the speech, it leads us to the design of speech enhancement algorithms based on the loudness model of the speech. The objective will be to enhance speech in the loudness domain instead of the spectral or time domains. In this way, we can derive some algorithms to improve the PESQ score and hence the speech quality. We also selected to include other quality measures that assess the noise level and the distortion level of the speech signal. The following chapters will investigate speech enhancement based on the statistical model and loudness model. The results will be evaluated with the P.862 (PESQ scores), segmental SNR and log likelihood ratio.

Chapter 3

Recursive Speech Enhancement Based on Laplacian-Gaussian Model

3.1 Introduction

Most speech enhancement algorithms assume that the speech and noise are independent zero mean Gaussian distributed random variables. This assumption has also been applied to the coefficients in the transformed domains. The Wiener filtering [31] and Kalman filtering have been used under such jointly Gaussian assumptions.

However, the statistical modeling of speech has been thoroughly discussed in [118], where it was shown that the speech coefficients in the transformed domains (Karhunen-Loève or Discrete Cosine transforms) are better modelled as zero mean Laplacian random variables. Therefore, the noisy speech signal is better represented as a mixture of Laplacian and Gaussian random variables. A Bayesian-based speech enhancement algorithms has been developed in [118]. The enhancement is achieved by estimating the clean speech components from the de-correlated noisy speech components with Minimum Mean Square

Error (MMSE) or Maximum Likelihood (ML) estimations. The estimation of the Laplacian factor is critical in noisy environments. To avoid the limitations caused by the error in the Laplacian factor estimation, a recursive speech enhancement algorithm is proposed next in this chapter.

The rest of this chapter is organized into five sections. First, the proposed speech enhancement system structure is given in Section 3.2. This speech enhancement, based on Laplacian-Gaussian model [118], is reviewed and evaluated in Section 3.3. The effects of the error in the Laplacian factor estimation on the previous speech enhancement algorithm are discussed in Section 3.4. The new recursive model is proposed in Section 3.5. Finally, conclusions are provided in Section 3.6.

3.2 Review of Speech Enhancement Employing Laplacian-Gaussian Model

In this section, a speech enhancement algorithm [48] [118] based on Laplacian-Gaussian Modelling is reviewed. In [47] [117] [118], a Laplacian model is proposed for the Karhunen-Loève transform (KLT) components of the speech signals. The noisy speech components are viewed as a Laplacian plus Gaussian mixture. The clean speech KLT components are then estimated from the mixture based on Bayesian estimation. Finally, the enhanced speech is reconstructed with the Inverse KLT of the clean speech KLT estimates.

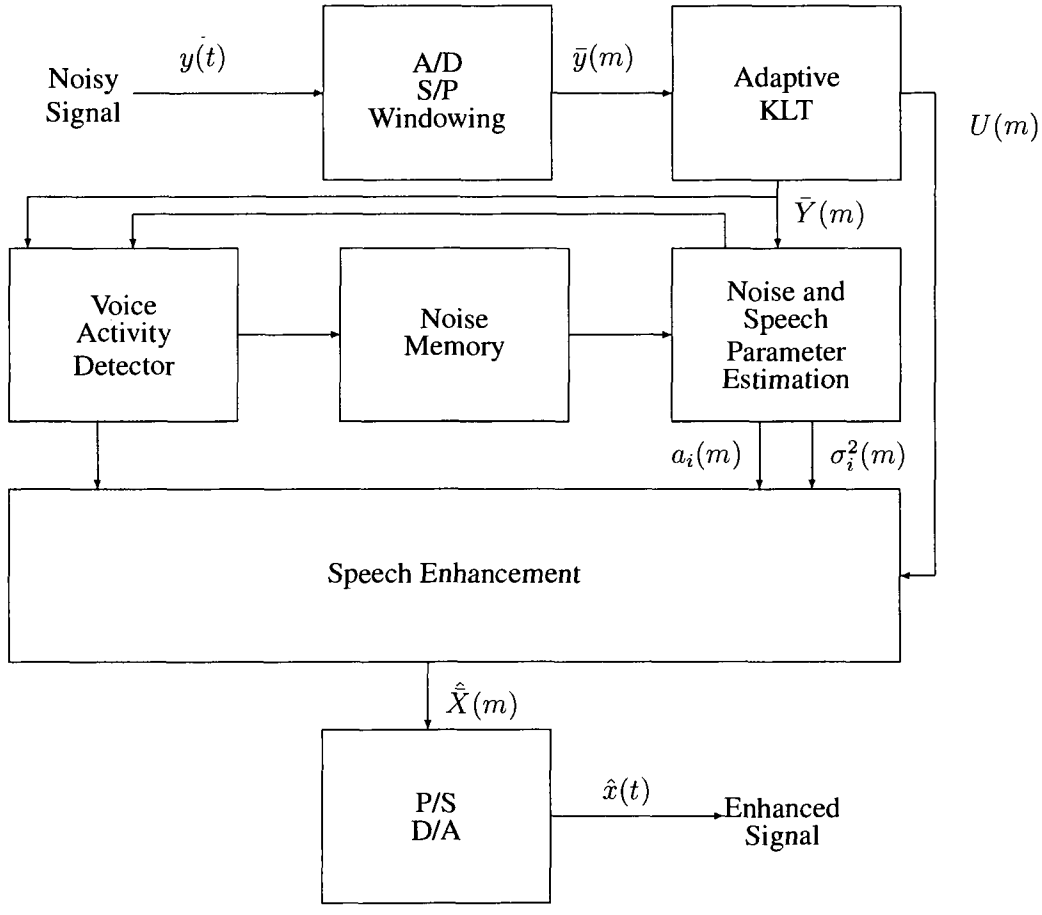

Figure 3.1: The block diagram of the statistical model-based speech enhancement system

3.2.1 Laplacian-Gaussian Model Based Speech Enhancement Algorithm

The speech enhancement algorithm in [48] is based on the assumption that the speech signal components in the KLT domain have Laplacian distributions as described in [33] [118]. The additive noise KLT components are assumed to be Gaussian distributed.

The speech enhancement approach is based on a Bayesian estimation of the KLT components of clean speech $X_i(m)$ from the noisy speech $Y_i(m)$ assuming that the speech Laplacian factor $a_i(m)$ and the noise variance $\sigma_i^2(m)$ are known. Both $a_i(m)$ and $\sigma_i^2(m)$ are estimated using Maximum Likelihood (ML) estimation method. The details will be given later.

Initialize :

$$d_i(0) = 0, \quad \beta = 0.973,$$
$$W(0) = [w_1(0)|w_2(0)|\ldots|w_K(0)] = I_K,$$

β is chosen to let the time constant of the following adaptive process to be 10 ms.

For each time step m do,

$$Y_1(m) = Y(m)$$
$$N_1(m) \leftarrow \text{Read from noise memory}$$

For i= 1, 2,. . ., K do

$$v_i(m) = w_i^T(m-1)Y_i(m)$$
$$u_i(m) = w_i^T(m-1)N_i(m)$$
$$d_i(m) = \beta d_i(m-1) + |v_i(m)|^2$$
$$E_i(m) = Y_i(m) - w_i(m-1)v_i(m)$$
$$w_i(m) = w_i(m-1) + E_i(m)\frac{v_i(m)}{d_i(m)}$$
$$Y_{i+1}(m) = Y_i(m) - w_i(m)v_i(m)$$
$$N_{i+1}(m) = N_i(m) - w_i(m)u_i(m)$$

end;

$$W(m) = [w_1(m)|w_2(m)|\ldots|w_K(m)]$$

end;

Table 3.1: KLT Tracking Algorithm

The block diagram of the speech enhancement system [118] is shown in Figure 3.1. The analog speech source is first passed through an analog to digital (A/D) converter. Then the digitalized signal is fed to a serial to parallel (S/P) converter, in which a rectangular window is used to obtain the noisy speech signal vector. This noisy speech vector will be noted as $\bar{y}(m)$.

The KLT of $\bar{y}(m)$ is denoted $\bar{Y}(m)$ with components $Y_i(m)$. The PASTd algorithm [112] is used here to implement the KLT as shown in Table 3.1. The transformation matrix is $U(m)$. $\bar{Y}(m)$, the KLT vector of the noisy speech is then used to estimate the noise and speech parameters that define the Laplacian factor of the speech component $a_i(m)$ and the variance of the noise $\widehat{\sigma_i^2}(m)$.

Given that the speech signal includes active intervals as well as noise intervals, a voice activity detector is deployed [46] [118]. The Voice Activity Detector separates the silent and active speech intervals from the noisy speech, $\bar{y}(m)$. The samples from the silent intervals are stored in a noise memory to estimate $\widehat{\sigma_i^2}$ based on the KLT transformation matrix $U(m)$ and the reference data from the noise memory.

Then, the clean speech components are estimated from the speech and noise parameters and the noisy speech $\bar{Y}(m)$. The details will be described later in this section. The enhanced speech vector is then derived as $\hat{\bar{X}}(m)$. After the Parallel to Serial (P/S) and Digital to Analog (D/A) converters, the enhanced time domain signal $\hat{x}(t)$ is obtained.

3.2.2 Speech Signal Modelling

It was shown that the KLT components of a clean speech signal are better described by a Laplacian distribution rather than a Gaussian distribution [33] [117] [118]. Assume that the KLT components $X_i(m)$ have zero-mean Laplacian distribution and are uncorrelated.

$f_{\mathbf{X}_i}$, the PDF of $X_i(m)$ is given by:

$$f_{\mathbf{X}_i}(X_i(m)) = \frac{1}{2a_i(m)} e^{-\frac{|X_i(m)|}{a_i(m)}}, \quad \forall i = 1, 2, \ldots, K, \tag{3.1}$$

where $a_i(m)$ is the i^{th} Laplacian factors for clean speech.

Assume the values of X_i during a length-M_S period $[(m - M_S + 1), m)]$ are known as

$$\tilde{X}_i(m) = [X_i(m), X_i(m-1), \ldots, X_i(m - M_S + 1)]^T. \tag{3.2}$$

To estimate the i^{th} Laplacian factor, $\hat{a}_i$, the ML Estimate [5] is given by:

$$\begin{aligned}
\hat{a}_i &= \arg\max_{a_i} \prod_{t=(m-M_S+1)}^{m} \frac{1}{2a_i} e^{-\frac{|X_i(t)|}{a_i}} \\
&= \arg\max_{a_i} \left(\frac{1}{2a_i}\right)^{M_S} e^{-\frac{1}{a_i}\sum_{t=(m-M_S+1)}^{m}|X_i(t)|} \\
&= \arg\max_{a_i} \left(M_S \log\frac{1}{2a_i} - \frac{1}{a_i} \sum_{t=(m-M_S+1)}^{m} |X_i(t)| \right).
\end{aligned} \tag{3.3}$$

Differentiating the above equation with respect to a_i and equating the derivative to zero

$$\begin{aligned}
\frac{\partial \hat{a}_i}{\partial a_i} &= \frac{\partial}{\partial a_i} \left(M_S \log\frac{1}{2a_i} - \frac{1}{a_i} \sum_{t=(m-M_S+1)}^{m} |X_i(t)| \right) \\
&= \left(\frac{-1}{2{a_i}^2}\right) \cdot 2a_i \cdot M_S - \left(\frac{-1}{{a_i}^2}\right) \sum_{t=(m-M_S+1)}^{m} |X_i(t)| = 0,
\end{aligned} \tag{3.4}$$

we get the ML estimate of a_i,

$$\hat{a}_i = \frac{\sum_{t=(m-M_S+1)}^{m} |X_i(t)|}{M_S}. \tag{3.5}$$

Equation (3.5) above shows that Laplacian factor a_i can be estimated using the moving average of $|X_i|$. In real-time process, we may use a recursive estimate of this average:

$$\hat{a}_i(m) = \beta_X \hat{a}_i(m-1) + (1 - \beta_X) |X_i(m)|. \tag{3.6}$$

where $\beta_X = \frac{M_X - 1}{M_X}$. (3.5) requires only two multiplications and one addition at each time step. In our simulations, β_X is chosen to set the time constant of above adaptive process to 10 ms because speech is generally considered to be stationary in 20-40 ms intervals.

Note that a_i in (3.5) is an estimate of the expected value of $|X_i(m)|$. Because we cannot access the real value of $X_i(m)$ during the process, we need to find a substitute for $|X_i(m)|$. Note that the added noise N is zero mean, independent of the speech signal and has a Gaussian distribution. So we may use $|Y_i(m)|$ as a substitute of $|X_i(m)|$ in (3.6), where $Y_i(m)$ is the noisy speech component corresponding to $X_i(m)$. This assumption is reasonable when the SNR is high.

3.2.3 Noise Signal Modelling

It is generally assumed that practical additive noise has a Gaussian PDF. Assume that the noise components along each eigenvector $p_{N_i}(N_i)$ has a Gaussian PDF:

$$p_{N_i}(N_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{N_i^2}{2\sigma_i^2}} \tag{3.7}$$

and that the successive samples of $N_i(m)$, during time interval between $(m - M_N + 1)$ and m, are independent and the variation of their variances are very small. In this case, the ML Estimation [5] of σ_i^2 is given by:

$$\begin{aligned}
\hat{\sigma}_i &= \arg\max_{\sigma_i^2} \prod_{t=(m-M_N+1)}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{N_i^2(t)}{2\sigma_i^2}} \\
&= \arg\max_{\sigma_i^2} \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \right)^{M_N} e^{-\frac{1}{2\sigma_i^2} \sum_{t=(m-M_N+1)}^{m} N_i^2(t)} \\
&= \arg\max_{\sigma_i^2} \left(M_N \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \right) - \frac{1}{2\sigma_i^2} \sum_{t=(m-M_N+1)}^{m} N_i^2(t) \right).
\end{aligned}$$

Differentiating the above equation with respect to σ_i^2 and setting the result to zero

$$\begin{aligned}
&\frac{\partial}{\partial(\sigma_i^2)}\left(M_N \cdot \left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right) - \frac{1}{2\sigma_i^2}\sum_{t=(m-M_N+1)}^{m} N_i^2(t)\right) \\
&= (\frac{-1}{\sigma_i^2}) \cdot \frac{1}{M_N} - (\frac{-1}{2\sigma_i^2}) \cdot \sum_{t=(m-M_N+1)}^{m} N_i^2(t) \\
&= 0, \qquad (3.8)
\end{aligned}$$

we get the ML estimate of σ_i^2,

$$\hat{\sigma}_i^2 = \frac{\sum_{t=(m-M_N+1)}^{m} N_i^2(t)}{M_N}. \qquad (3.9)$$

Thus, the noise variance σ_i^2 can be estimated as the moving average of N_i^2. Again, this average can be estimated recursively as:

$$\widehat{\sigma_i^2}(m) = \beta_N \widehat{\sigma_i^2}(m-1) + (1-\beta_N)|N_i(m)|^2. \qquad (3.10)$$

where $\beta_N = \frac{M_N - 1}{M_N}$. In our simulations, β_N is chosen to set the time constant of the above adaptive process to be 0.5 second, in which the variation of noise variance can be assumed to be negligible.

3.2.4 Estimation of Clean Speech Components

As the KLT components can be assumed to be uncorrelated, we first consider just one component. We drop the time index m and component index i for simplicity for the remaining of this subsection. The main objective of this part is to estimate the clean speech component X from the noisy speech component Y in the KLT domain when the speech is active. In this case

$$Y = X + N, \qquad (3.11)$$

where

$$f_{\mathbf{X}}(X) = \frac{1}{2a} e^{-\frac{|X|}{a}}, \tag{3.12}$$

and

$$f_{\mathbf{N}}(N) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{N^2}{2\sigma^2}}. \tag{3.13}$$

Assuming that speech component X and noise component N are independent, the joint distribution [4] of X and Y is given by

$$f_{\mathbf{X},\mathbf{Y}}(X,Y) = \frac{1}{2a\sqrt{2\pi\sigma^2}} e^{-\frac{|X|}{a} - \frac{|Y-X|^2}{2\sigma^2}}. \tag{3.14}$$

To estimate the speech component X, we use the MMSE (Minimum Mean Square Error) Estimation or ML (Maximum Likelihood) Estimation [5] described below.

The MMSE estimate of X [118] can be derived as a function with three inputs Y, σ^2 and a as following:

$$\begin{aligned}
\text{MMSE}: \quad \hat{X} &\triangleq E\{X|Y\} \\
&= \int X f_{\mathbf{X}|\mathbf{Y}}(X|Y) dX \\
&= \frac{\left(Y + \frac{\sigma^2}{a}\right) e^{\frac{\sigma^2}{2a^2} + \frac{Y}{a}} \operatorname{erfc}\left(\frac{aY+\sigma^2}{\sqrt{2\sigma^2}a}\right) + \left(Y - \frac{\sigma^2}{a}\right) e^{\frac{\sigma^2}{2a^2} - \frac{Y}{a}} \operatorname{erfc}\left(\frac{-aY+\sigma^2}{\sqrt{2\sigma^2}a}\right)}{e^{\frac{\sigma^2}{2a^2} + \frac{Y}{a}} \operatorname{erfc}\left(\frac{aY+\sigma^2}{\sqrt{2\sigma^2}a}\right) + e^{\frac{\sigma^2}{2a^2} - \frac{Y}{a}} \operatorname{erfc}\left(\frac{-aY+\sigma^2}{\sqrt{2\sigma^2}a}\right)},
\end{aligned} \tag{3.15}$$

where $\operatorname{erfc}(x)$ is the complementary error function:

$$\operatorname{erfc}(x) = \int_x^\infty \exp(-t^2) dt. \tag{3.16}$$

If we let $\xi = \frac{Y}{a}$ and $\psi = \frac{\sigma^2}{a^2}$, the above equation can be simplified to

$$\begin{aligned}\textbf{MMSE}: \quad \hat{X} &\triangleq E\{X|Y\} \\ &= \int X f_{\mathbf{X}|\mathbf{Y}}(X|Y)dX \\ &= a\left[\frac{(\xi+\psi)\,e^{\frac{\psi}{2}+\xi}\mathrm{erfc}\left(\frac{\xi+\psi}{\sqrt{2\psi}}\right)+(\xi-\psi)\,e^{\frac{\psi}{2}-\xi}\mathrm{erfc}\left(\frac{-\xi+\psi}{\sqrt{2\psi}}\right)}{e^{\frac{\psi}{2}+\xi}\mathrm{erfc}\left(\frac{\xi+\psi}{\sqrt{2\psi}}\right)+e^{\frac{\psi}{2}-\xi}\mathrm{erfc}\left(\frac{-\xi+\psi}{\sqrt{2\psi}}\right)}\right].\end{aligned} \tag{3.17}$$

While ML Estimation of X is [118]:

$$\begin{aligned}\textbf{ML}: \quad \hat{X} &\triangleq \arg\max_X f_{\mathbf{Y}|\mathbf{X}}(Y|X) \\ &= \arg\max_X f_{\mathbf{X},\mathbf{Y}}(X,Y) \\ &= \arg\min_X \left(\frac{|X|}{a}+\frac{|Y-X|^2}{2\sigma^2}\right) \\ &= \begin{cases} Y-\frac{\sigma^2}{a} & Y \geq \frac{\sigma^2}{a} \\ 0 & |Y| \leq \frac{\sigma^2}{a} \\ Y+\frac{\sigma^2}{a} & Y \leq -\frac{\sigma^2}{a} \end{cases}\end{aligned} \tag{3.18}$$

The relationship between these two estimates is shown in Figure 3.2. We find that both will be similar for $|Y| \geq 2\frac{\sigma^2}{a}$, i.e., high instantaneous SNR. At low instantaneous SNRs, the ML tends to be more aggressive in noise reduction and possibly result in more distortion to the speech.

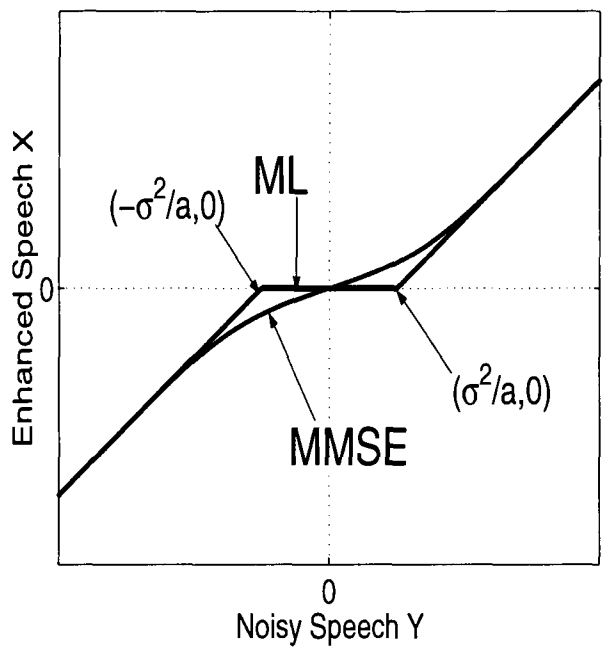

Figure 3.2: Comparison of the MMSE and ML approach

3.3 Experimental Results

3.3.1 KLT-based Speech Enhancement [118]

The test results while applying KLT transformation are shown in Table 3.2. The speech algorithm based on the Wiener filter [94] will be used here for comparison. The frame length used for testing is 80 samples which is 10 ms of speech signals. The KLT is derived base on the PASTd algorithm [112]. We estimated the Laplacian factor with the noisy speech (SE in [118]) and the clean speech signals (Ideal SE in [118]).

We can see that the KLT approach with MMSE improves the speech quality (as measured by the PESQ) by 0.19-0.28 for female and 0.01-0.11 for male. The improvement is

Speech Type	SNR	Input Signal	PESQ Improvement				
			SE in [118]		Ideal SE in [118]		SE with Wiener Filter
			MMSE	ML	MMSE	ML	
Female	0dB	1.15	+0.19	+0.21	-0.06	-0.46	+0.20
	5dB	1.45	+0.28	+0.33	+0.16	-0.13	+0.33
	10dB	1.82	+0.28	+0.32	+ 0.20	-0.06	+0.30
	15dB	2.24	+0.23	+0.21	+0.18	+0.02	+0.20
	20dB	2.63	+0.18	+0.14	+0.17	+0.05	+0.14
Male	0dB	1.40	+0.01	-0.20	-0.35	-1.04	-0.27
	5dB	1.59	+0.02	-0.04	-0.09	-0.77	-0.05
	10dB	1.84	+0.07	+0.05	-0.01	-0.33	+0.03
	15dB	2.14	+0.10	+0.05	+0.06	-0.08	+0.06
	20dB	2.49	+0.11	+0.11	+0.09	+0.08	+0.11

Table 3.2: Comparison of PESQ improvements of input and enhanced signals for KLT-based approaches

relatively lower for very low SNRs. We can see that even though the SNRs were improved as shown in [118], the algorithm did not improve the PESQ values for some of the male speech cases. This is due to the fact that the KLT is a data-dependent transform and the data is noise-corrupted. So the KLT transformation matrix itself will be effected by the additive noise and the time-variation of the speech signal, especially for the low SNR case. When the additive noise energy is comparable to the speech signal, the performance of this enhancement algorithm will be degraded. With the ideal estimation of the Laplacian factor, the enhanced speech contains less noise. At the same time, the enhanced speech is more likely to be distorted. This distortion becomes even more problematic with KLT since the KLT is data-dependent and time varying. This leads us to test the same algorithm with data-independent DCT transform.

3.3.2 DCT-based SE

The speech enhancement algorithm described in the previous section can be modified by using the DCT instead of the KLT. It is computationally-efficient and it can whiten an autoregressive signal (e.g. speech) almost as well as KLT, if the data size is large enough [62]. The major advantage of DCT over KLT is that DCT has the data-independent transformation basis, while the KLT basis is data-dependent and time-varying with the non-stationary speech signals. This reduces the complexity of calculating the transforming matrix significantly and also makes the transformation independent of the noise. The test results with DCT transformation are shown in Table 3.3 and 3.4. The speech enhancement algorithm based on the Wiener filter [94] will be used here for comparison.

Input SNR	Input PESQ	PESQ Improvement				
		SE in [118]		Ideal SE in [118]		SE with Wiener filter
		ML	MMSE	ML	MMSE	
0dB	1.15	+0.31	+0.22	+0.58	+0.73	+0.42
5dB	1.45	+0.41	+0.29	+0.50	+0.85	+0.49
10dB	1.82	+0.40	+0.31	+0.48	+0.75	+0.47
15dB	2.24	+0.36	+0.29	+0.52	+0.62	+0.41
20dB	2.63	+0.35	+0.29	+0.51	+0.47	+0.39

Table 3.3: Comparison of PESQ improvements for DCT-based approaches (Female speech)

Input SNR	Input PESQ	PESQ Improvement				
		SE in [118]		Ideal SE in [118]		SE with Wiener filter
		ML	MMSE	ML	MMSE	
0dB	1.40	+0.14	+0.10	+0.05	+0.17	+0.17
5dB	1.59	+0.23	+0.14	+0.30	+0.67	+0.29
10dB	1.84	+0.34	+0.23	+0.40	+0.71	+0.41
15dB	2.14	+0.37	+0.27	+0.47	+0.64	+0.42
20dB	2.49	+0.35	+0.27	+0.51	+0.52	+0.39

Table 3.4: Comparison of PESQ improvements for DCT-based approaches (Male speech)

From Table 3.3, we can clearly see that this speech enhancement algorithm improves the

speech quality for each case compared to the unprocessed noisy signal. By comparing with Table 3.2, the DCT leads to better performance of enhancement than the KLT. However this DCT-based algorithm does not outperform the Wiener filter approach. The main reason for this result is the error in the estimation of the Laplacian factor resulting from the use of the noisy signal instead of the actual clean speech. To isolate the effect of the Laplacian factor estimation, the 'Ideal SE' columns of the Tables 3.3 and 3.4 show the results of the same experiment using the actual Laplacian factor obtained from the clean signals. We can see that the use of an exact Laplacian factor with MMSE estimation clearly outperforms the Wiener filter approach. And the ML estimation also outperforms the Wiener filter approach in high SNR cases. Thus in the next section, we will try to improve the estimation of the Laplacian factor in order to improve the overall performance of this speech enhancement algorithm.

Table 3.5 shows an example of the Mean Squared Error (MSE) of the noisy and enhanced speech compared to the clean speech. Since the enhanced signal generally has smaller MSE value, it can provide a better reference for the Laplacian factor estimation than the noisy signal itself, hence improving the performance of the speech enhancement. This leads to the new contribution of this chapter – recursive estimation of the Laplacian factor. This approach will be discussed later in this chapter.

Input SNR	Input Signal	SE in [118]		Ideal SE in [118]		SE with Wiener filter
		ML	MMSE	ML	MMSE	
0dB	0.14	0.07	0.08	0.05	0.06	0.07
5dB	0.08	0.05	0.05	0.04	0.04	0.05
10dB	0.04	0.03	0.03	0.03	0.03	0.03
15dB	0.03	0.02	0.02	0.02	0.02	0.02
20dB	0.01	0.02	0.02	0.02	0.02	0.02

Table 3.5: Comparison of MSE values for different approaches

3.3.3 Complexity of the algorithm

The computational complexity of the whole system is the combination of the cost of the transformation and the enhancement. For the transformation, the computational complexity of DCT and IDCT is of the order of $K\log_2(\mathrm{K})$, where the K is the length of the input vector. The computational complexity of the KLT is of the order $4Kr + O(r)$, where r is the number of eigenvectors of the covariance matrix of clean speech signal X [112].

The complexity of the speech enhancement algorithm is described next. We only include computations other than adding or subtraction. From (3.18), the ML based algorithm only needs one additional division per component. However, for the MMSE based algorithm as in (3.17), we need 6 multiplications, 4 divisions, 2 square roots, 2 exponential, and 2 'erfc' functions for each component. It should be noted that for KLT, fewer components are generally needed due to the reduced correlation between the components.

3.4 Relationship between Laplacian Factor Estimation and Speech Enhancement

The variance of the Gaussian factor of the noise signal can be estimated within the silent intervals (noise only) of the speech signal. On the other hand, the speech signal is always corrupted by the noise signal. Hence the Laplacian factor of the speech components cannot be estimated from the speech signal directly. In the previous section, the Laplacian factor was estimated with noisy speech signals. This becomes problematic for the enhanced signals, especially when the SNR is low.

Assuming we are still considering one DCT component, Figure 3.3 shows the effect of the estimation error of the parameters on the enhanced speech. The x-axis is the noisy

speech Y and the y-axis is the enhanced speech X, both in the transformed domain. The speech enhancement algorithms applied here are ML and MMSE approaches presented in the previous Section 3.2. As we can see from Equation (3.17), this enhancement algorithm was mainly controlled by the value of σ^2/a, which is denoted by Th in Figure 3.3. Assume that the Th_1 is the ideal value from the clean speech signal and Th_2 is the estimated value from the noisy speech signal. The a is over-estimated by using Y instead of X due to the extra N term. So the Th_2 is smaller than Th_1 under such circumstances.

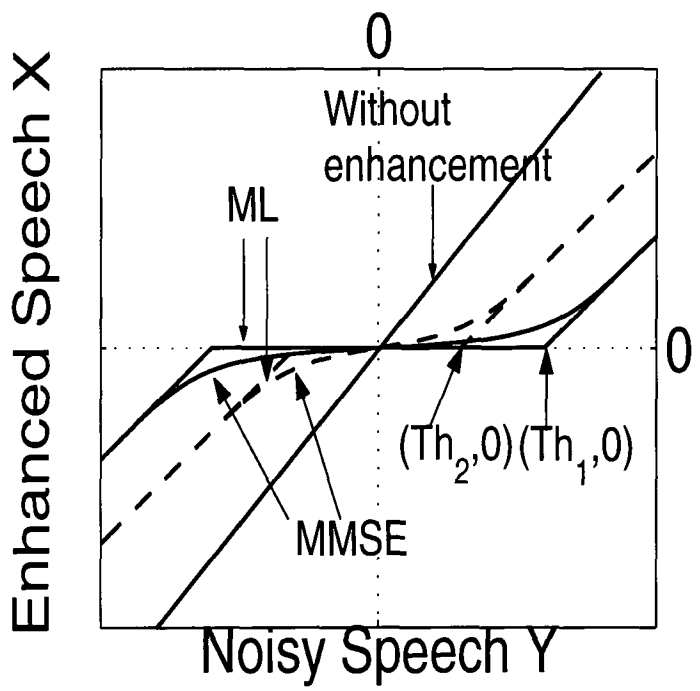

Figure 3.3: The effect of Laplacian factor estimation error

In Figure 3.3, the solid line through the origin represents the no-enhancement case. The two other solid curves present the case of speech enhancement controlled by the ideal factor Th_1. The two dashed curves are controlled by the estimated factor Th_2. For the proposed Laplacian estimation, Th_2 is generally smaller than Th_1.

Clearly, using Th_1 will cut more of the noisy speech component than using Th_2. The ideal approach (with Th_1) will be the best estimation of the Laplacian factor, in the mathematical ML or MMSE sense. However, it should be noted that the ideal case may not necessarily correspond to the best quality of enhancement, given that the quality of the enhanced signal is determined by not only the level of noise removal, but also the distortion of the enhanced speech from the clean speech.

3.5 Proposed Recursive Laplacian Factor Estimation and Speech Enhancement

As we have shown in the previous section, the enhanced speech signal will have a smaller MSE compared to the clean speech signal than the noisy signal. In this section, a new improvement of the previous speech enhancement algorithm will be introduced. We propose to use the enhanced speech signal as reference signal for the estimation of the Laplacian factor, which is then used to further enhance the original signal. Since the enhanced signal is expected to have a better estimate of the Laplacian factor, this improved estimate is expected to result in a further improved enhanced signal.

The recursive estimation is done as follows: the output signal of the previous iteration is used as the reference signal to estimate the Laplacian factor for the next iteration. As shown, the ML estimation is computational more efficient, while the MMSE estimation is more accurate in the MMSE sense. Thus the enhanced speech from ML estimation can be used for ML estimation at the subsequent iteration to maintain the computationally efficiency and remove the noise faster. Alternatively, enhanced speech from MMSE estimation approaches is used for MMSE estimation in subsequent iterations to achieve lower MSE.

A combination of these approaches is also possible: ML estimation is applied for the first several iterations for quick removal of the noise and computationally efficiency, followed by an iteration of MMSE estimation to achieve lower MSE.

The performance based on this recursive estimation of the Laplacian factor is compared with the approach based on Wiener filter [94] and the results using ideal Laplacian factor estimation, which is directly obtained from the clean speech signal. These results are given in Tables 3.6 to 3.11 for different number of iterations in the recursive estimation of the Laplacian factor.

Tables 3.6 to 3.11 also provide the PESQ values of the input speech (noisy speech) and the PESQ score improvements for each enhanced speech are listed. All the improvements are compared to the scores of input speech. We also give the score improvements with Wiener filter approach for comparison. To complete the quality assessment of the enhanced speech, Tables 3.12 to 3.17 give the segmental SNR values of the noisy and enhanced speeches and Tables 3.18 to 3.23 give the Log Likelihood Ratio (LLR) values of the noisy and enhanced speeches. In all Tables 3.6 to 3.23, the "Ideal SE" refers to results using the clean speech for the Laplacian factor estimation. The column "ML after ML(s)" and "MMSE after MMSE(s)" refer to the case when all the iterations are based on the ML and MMSE approaches, respectively. The column "MMSE after ML(s)" refers to the case when all the iterations are ML except for the last iteration which is MMSE.

In Tables 3.24 and 3.25, we compare all the quality measures as well as the remaining energy in the enhanced speech after each iteration for the female speech 3. The remaining energy after each iteration is compared to the energy of the clean speech signal. When the remaining energy is more than 100% of the energy of clean speech, the parameter a tends to be over-estimated, and vice versa.

Input SNR	Iteration	Input Signal PESQ	PESQ Improvement			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	1.36	+0.20		+0.16	+0.21
	2		+0.30	+0.27	+0.25	
	3		+0.32	+0.36	+0.32	
	4		+0.21	+0.40	+0.39	
	5		+0.04	+0.36	+0.38	
	6		-0.06	+0.23	+0.27	
Ideal SE			+0.41		+0.69	
5dB	1	1.64	+0.29		+0.21	+0.31
	2		+0.43	+0.38	+0.35	
	3		+0.38	+0.46	+0.44	
	4		+0.26	+0.48	+0.48	
	5		+0.20	+0.37	+0.42	
	6		+0.13	+0.26	+0.29	
Ideal SE			+0.37		+0.69	
10dB	1	1.93	+0.35		+0.26	+0.37
	2		+0.46	+0.42	+0.40	
	3		+0.47	+0.49	+0.47	
	4		+0.43	+0.51	+0.50	
	5		+0.39	+0.47	+0.48	
	6		+0.36	+0.44	+0.45	
Ideal SE			+0.55		+0.68	
15dB	1	2.32	+0.37		+0.30	+0.37
	2		+0.47	+0.42	+0.41	
	3		+0.48	+0.49	+0.46	
	4		+0.44	+0.49	+0.48	
	5		+0.38	+0.47	+0.48	
	6		+0.33	+0.43	+0.45	
Ideal SE			+0.55		+0.63	
20dB	1	2.67	+0.35		+0.27	+0.34
	2		+0.45	+0.39	+0.37	
	3		+0.49	+0.46	+0.43	
	4		+0.46	+0.47	+0.46	
	5		+0.44	+0.46	+0.46	
	6		+0.43	+0.45	+0.45	
Ideal SE			+0.53		+0.56	

Table 3.6: Evaluation of recursive speech enhancement approaches: PESQ (Female Speech 1)

Input SNR	Iteration	Input Signal PESQ	PESQ Improvement			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	1.60	+0.20		+0.17	+0.22
	2		+0.27	+0.28	+0.26	
	3		+0.25	+0.34	+0.32	
	4		+0.10	+0.35	+0.34	
	5		+0.05	+0.15	+0.25	
	6		-0.16	+0.00	+0.07	
Ideal SE			+0.32		+0.50	
5dB	1	1.87	+0.27		+0.21	+0.28
	2		+0.31	+0.32	+0.30	
	3		+0.28	+0.35	+0.34	
	4		+0.03	+0.31	+0.34	
	5		-0.20	+0.13	+0.19	
	6		-0.27	-0.05	+0.00	
Ideal SE			+0.37		+0.49	
10dB	1	2.17	+0.31		+0.22	+0.32
	2		+0.40	+0.37	+0.34	
	3		+0.33	+0.43	+0.41	
	4		+0.04	+0.38	+0.42	
	5		-0.22	+0.15	+0.25	
	6		-0.34	-0.07	+0.00	
Ideal SE			+0.43		+0.45	
15dB	1	2.44	+0.30		+0.23	+0.31
	2		+0.40	+0.37	+0.34	
	3		+0.33	+0.43	+0.41	
	4		+0.20	+0.35	+0.39	
	5		+0.11	+0.23	+0.28	
	6		+0.01	+0.13	+0.19	
Ideal SE			+0.41		+0.46	
20dB	1	2.73	+0.27		+0.23	+0.28
	2		+0.42	+0.36	+0.34	
	3		+0.37	+0.41	+0.39	
	4		+0.28	+0.35	+0.38	
	5		+0.23	+0.29	+0.30	
	6		+0.20	+0.25	+0.25	
Ideal SE			+0.33		+0.42	

Table 3.7: Evaluation of recursive speech enhancement approaches: PESQ (Female Speech 2)

Input SNR	Iteration	Input Signal PESQ	PESQ Improvement			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	1.25	+0.34		+0.26	+0.35
	2		+0.47	+0.44	+0.41	
	3		+0.48	+0.54	+0.52	
	4		+0.35	+0.58	+0.59	
	5		+0.16	+0.49	+0.53	
	6		+0.06	+0.27	+0.33	
Ideal SE			+0.37		+0.64	
5dB	1	1.60	+0.39		+0.30	+0.41
	2		+0.53	+0.49	+0.46	
	3		+0.47	+0.57	+0.54	
	4		+0.27	+0.56	+0.57	
	5		+0.10	+0.40	+0.51	
	6		-0.02	+0.19	+0.25	
Ideal SE			+0.41		+0.70	
10dB	1	1.96	+0.40		+0.29	+0.41
	2		+0.54	+0.49	+0.46	
	3		+0.51	+0.55	+0.54	
	4		+0.32	+0.52	+0.54	
	5		+0.17	+0.45	+0.49	
	6		+0.11	+0.33	+0.39	
Ideal SE			+0.48		+0.69	
15dB	1	2.33	+0.43		+0.33	+0.43
	2		+0.52	+0.47	+0.44	
	3		+0.55	+0.53	+0.51	
	4		+0.54	+0.55	+0.54	
	5		+0.51	+0.55	+0.54	
	6		+0.49	+0.53	+0.54	
Ideal SE			+0.63		+0.70	
20dB	1	2.70	+0.39		+0.30	+0.38
	2		+0.49	+0.40	+0.38	
	3		+0.51	+0.46	+0.43	
	4		+0.51	+0.46	+0.44	
	5		+0.50	+0.45	+0.44	
	6		+0.48	+0.44	+0.43	
Ideal SE			+0.58		+0.62	

Table 3.8: Evaluation of recursive speech enhancement approaches: PESQ (Female Speech 3)

Input SNR	Iteration	Input Signal PESQ	PESQ Improvement			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	1.39	+0.12		+0.08	+0.14
	2		+0.25	+0.20	+0.17	
	3		+0.32	+0.32	+0.28	
	4		+0.23	+0.41	+0.40	
	5		+0.14	+0.35	+0.39	
	6		+0.00	+0.24	+0.27	
Ideal SE			+0.27		+0.70	
5dB	1	1.57	+0.26		+0.16	+0.27
	2		+0.43	+0.35	+0.31	
	3		+0.47	+0.48	+0.44	
	4		+0.36	+0.54	+0.52	
	5		+0.24	+0.51	+0.52	
	6		+0.17	+0.42	+0.46	
Ideal SE			+0.38		+0.71	
10dB	1	1.83	+0.36		+0.25	+0.37
	2		+0.52	+0.45	+0.42	
	3		+0.49	+0.56	+0.53	
	4		+0.36	+0.56	+0.57	
	5		+0.22	+0.43	+0.49	
	6		+0.14	+0.32	+0.35	
Ideal SE			+0.47		+0.68	
15dB	1	2.15	+0.37		+0.30	+0.39
	2		+0.49	+0.46	+0.43	
	3		+0.47	+0.52	+0.51	
	4		+0.39	+0.51	+0.53	
	5		+0.29	+0.46	+0.50	
	6		+0.25	+0.40	+0.42	
Ideal SE			+0.50		+0.63	
20dB	1	2.49	+0.38		+0.31	+0.39
	2		+0.50	+0.48	+0.45	
	3		+0.48	+0.52	+0.52	
	4		+0.44	+0.51	+0.52	
	5		+0.38	+0.48	+0.49	
	6		+0.32	+0.42	+0.46	
Ideal SE			+0.46		+0.53	

Table 3.9: Evaluation of recursive speech enhancement approaches: PESQ (Male Speech 1)

Input SNR	Iteration	Input Signal PESQ	PESQ Improvement			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	1.63	+0.23		+0.17	+0.24
	2		+0.34	+0.29	+0.27	
	3		+0.42	+0.41	+0.37	
	4		+0.32	+0.49	+0.47	
	5		+0.15	+0.43	+0.46	
	6		+0.09	+0.34	+0.42	
Ideal SE			+0.33		+0.75	
5dB	1	1.95	+0.28		+0.21	+0.28
	2		+0.37	+0.36	+0.33	
	3		+0.32	+0.45	+0.41	
	4		+0.27	+0.41	+0.41	
	5		+0.07	+0.33	+0.37	
	6		+0.06	+0.18	+0.24	
Ideal SE			+0.29		+0.68	
10dB	1	2.29	+0.35		+0.28	+0.37
	2		+0.42	+0.41	+0.39	
	3		+0.36	+0.44	+0.43	
	4		+0.28	+0.43	+0.45	
	5		+0.14	+0.35	+0.40	
	6		-0.00	+0.25	+0.29	
Ideal SE			+0.41		+0.63	
15dB	1	2.68	+0.36		+0.32	+0.38
	2		+0.39	+0.40	+0.40	
	3		+0.36	+0.42	+0.41	
	4		+0.27	+0.41	+0.41	
	5		+0.19	+0.35	+0.38	
	6		+0.14	+0.30	+0.33	
Ideal SE			+0.37		+0.57	
20dB	1	3.11	+0.21		+0.22	+0.18
	2		+0.27	+0.25	+0.24	
	3		+0.19	+0.25	+0.25	
	4		+0.13	+0.23	+0.24	
	5		+0.10	+0.12	+0.22	
	6		+0.07	+0.10	+0.11	
Ideal SE			+0.17		+0.36	

Table 3.10: Evaluation of recursive speech enhancement approaches: PESQ (Male Speech 2)

Input SNR	Iteration	Input Signal PESQ	PESQ Improvement			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	1.35	+0.32		+0.240	+0.34
	2		+0.45	+0.42	+0.39	
	3		+0.46	+0.55	+0.49	
	4		+0.22	+0.59	+0.56	
	5		+0.14	+0.49	+0.46	
	6		+0.08	+0.31	+0.36	
Ideal SE			+0.47		+0.92	
5dB	1	1.70	+0.36		+0.28	+0.38
	2		+0.50	+0.47	+0.43	
	3		+0.47	+0.59	+0.56	
	4		+0.36	+0.56	+0.59	
	5		+0.25	+0.46	+0.52	
	6		+0.12	+0.28	+0.30	
Ideal SE			+0.50		+0.81	
10dB	1	2.06	+0.40		+0.31	+0.42
	2		+0.50	+0.49	+0.44	
	3		+0.42	+0.54	+0.53	
	4		+0.26	+0.48	+0.52	
	5		+0.10	+0.36	+0.40	
	6		+0.06	+0.24	+0.27	
Ideal SE			+0.52		+0.71	
15dB	1	2.47	+0.44		+0.34	+0.42
	2		+0.52	+0.48	+0.46	
	3		+0.52	+0.53	+0.51	
	4		+0.44	+0.52	+0.52	
	5		+0.34	+0.44	+0.48	
	6		+0.30	+0.38	+0.41	
Ideal SE			+0.49		+0.60	
20dB	1	2.84	+0.37		+0.29	+0.34
	2		+0.41	+0.40	+0.38	
	3		+0.40	+0.41	+0.40	
	4		+0.37	+0.39	+0.40	
	5		+0.35	+0.38	+0.38	
	6		+0.32	+0.36	+0.37	
Ideal SE			+0.41		+0.47	

Table 3.11: Evaluation of recursive speech enhancement approaches: PESQ (Male Speech 3)

Input SNR	Iteration	Input Signal SegSNR	Segmental SNR			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	-6.56	-1.97		-2.98	-1.73
	2		-0.01	-0.72	-1.48	
	3		1.08	0.76	-0.60	
	4		1.69	1.61	-0.02	
	5		2.11	2.15	0.46	
	6		2.32	2.44	0.80	
Ideal SE			2.85		1.53	
5dB	1	-1.56	2.36		1.61	2.56
	2		3.81	3.38	2.83	
	3		4.51	4.41	3.49	
	4		4.92	4.99	3.90	
	5		5.18	5.33	4.22	
	6		5.32	5.51	4.45	
Ideal SE			6.09		5.24	
10dB	1	3.44	6.61		6.14	6.78
	2		7.48	7.35	7.01	
	3		7.88	7.97	7.43	
	4		8.07	8.27	7.68	
	5		8.21	8.47	7.86	
	6		8.28	8.56	7.99	
Ideal SE			9.15		8.77	
15dB	1	8.44	10.86		10.61	10.97
	2		11.33	11.39	11.19	
	3		11.49	11.71	11.43	
	4		11.58	11.86	11.57	
	5		11.62	11.92	11.66	
	6		11.64	11.95	11.71	
Ideal SE			12.46		12.36	
20dB	1	13.44	15.07		15.03	15.17
	2		15.25	15.44	15.35	
	3		15.30	15.58	15.47	
	4		15.27	15.62	15.53	
	5		15.25	15.62	15.56	
	6		15.27	15.63	15.60	
Ideal SE			15.67		15.74	

Table 3.12: Evaluation of recursive speech enhancement approaches: Segmental SNR (Female Speech 1)

Input SNR	Iteration	Input Signal SegSNR	Segmental SNR			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	-11.78	-7.24		-8.25	-7.01
	2		-5.10	-5.93	-6.72	
	3		-3.71	-4.25	-5.82	
	4		-2.59	-2.97	-5.11	
	5		-1.74	-2.01	-4.55	
	6		-1.15	-1.42	-4.12	
Ideal SE			-1.57		-3.38	
5dB	1	-6.78	-2.77		-3.57	-2.57
	2		-1.01	-1.60	-2.23	
	3		0.11	-0.25	-1.48	
	4		0.92	0.69	-1.00	
	5		1.58	1.43	-0.62	
	6		2.00	1.94	-0.35	
Ideal SE			1.69		0.43	
10dB	1	-1.78	1.73		1.13	1.90
	2		2.90	2.58	2.12	
	3		3.58	3.46	2.62	
	4		4.09	4.05	2.99	
	5		4.43	4.47	3.23	
	6		4.76	4.81	3.43	
Ideal SE			5.07		4.27	
15dB	1	3.22	6.02		5.63	6.15
	2		6.87	6.70	6.37	
	3		7.35	7.35	6.74	
	4		7.72	7.77	6.99	
	5		7.93	8.02	7.18	
	6		8.08	8.20	7.33	
Ideal SE			8.54		7.95	
20dB	1	8.22	10.35		10.14	10.43
	2		10.83	10.82	10.59	
	3		11.09	11.20	10.82	
	4		11.28	11.46	10.98	
	5		11.39	11.61	11.10	
	6		11.47	11.70	11.18	
Ideal SE			11.69		11.36	

Table 3.13: Evaluation of recursive speech enhancement approaches: Segmental SNR (Female Speech 2)

Input SNR	Iteration	Input Signal SegSNR	Segmental SNR			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	-6.76	-2.58		-3.41	-2.36
	2		-0.88	-1.45	-2.09	
	3		0.15	-0.12	-1.32	
	4		0.82	0.72	-0.83	
	5		1.39	1.33	-0.46	
	6		1.68	1.74	-0.17	
Ideal SE			1.95		0.80	
5dB	1	-1.76	1.93		1.28	2.11
	2		3.29	2.91	2.41	
	3		3.99	3.88	2.97	
	4		4.36	4.42	3.29	
	5		4.67	4.77	3.55	
	6		4.85	5.01	3.79	
Ideal SE			5.40		4.68	
10dB	1	3.24	6.38		5.90	6.51
	2		7.29	7.09	6.73	
	3		7.75	7.72	7.10	
	4		8.05	8.10	7.33	
	5		8.22	8.34	7.51	
	6		8.32	8.45	7.59	
Ideal SE			8.96		8.53	
15dB	1	8.24	10.63		10.39	10.73
	2		11.08	11.12	10.93	
	3		11.27	11.45	11.15	
	4		11.36	11.58	11.25	
	5		11.47	11.69	11.34	
	6		11.51	11.75	11.39	
Ideal SE			12.22		12.09	
20dB	1	13.24	15.01		14.90	15.05
	2		15.26	15.34	15.23	
	3		15.33	15.54	15.39	
	4		15.33	15.58	15.42	
	5		15.35	15.62	15.47	
	6		15.35	15.63	15.49	
Ideal SE			15.76		15.74	

Table 3.14: Evaluation of recursive speech enhancement approaches: Segmental SNR (Female Speech 3)

Input SNR	Iteration	Input Signal SegSNR	Segmental SNR			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	-6.30	-1.74		-2.75	-1.48
	2		0.25	-0.48	-1.26	
	3		1.44	1.06	-0.36	
	4		2.26	2.10	0.26	
	5		2.83	2.81	0.77	
	6		3.21	3.25	1.18	
Ideal SE			3.37		1.87	
5dB	1	-1.30	2.86		1.98	3.10
	2		4.59	4.00	3.32	
	3		5.53	5.28	4.08	
	4		6.14	6.07	4.61	
	5		6.48	6.55	5.02	
	6		6.71	6.83	5.34	
Ideal SE			6.77		5.66	
10dB	1	3.70	7.23		6.59	7.41
	2		8.48	8.13	7.66	
	3		9.17	9.06	8.24	
	4		9.51	9.55	8.57	
	5		9.70	9.83	8.82	
	6		9.89	10.03	9.02	
Ideal SE			9.98		9.06	
15dB	1	8.70	11.59		11.16	11.73
	2		12.36	12.24	11.92	
	3		12.74	12.82	12.33	
	4		12.93	13.10	12.53	
	5		13.04	13.26	12.68	
	6		13.11	13.34	12.80	
Ideal SE			13.32		13.04	
20dB	1	13.70	15.92		15.70	16.02
	2		16.30	16.38	16.20	
	3		16.43	16.65	16.41	
	4		16.46	16.74	16.51	
	5		16.48	16.78	16.59	
	6		16.51	16.82	16.63	
Ideal SE			16.67		16.61	

Table 3.15: Evaluation of recursive speech enhancement approaches: Segmental SNR (Male Speech 1)

Input SNR	Iteration	Input Signal SegSNR	Segmental SNR			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	-5.94	-2.12		-2.80	-1.94
	2		-0.57	-1.04	-1.60	
	3		0.23	0.03	-1.01	
	4		0.92	0.77	-0.56	
	5		1.55	1.37	-0.30	
	6		1.74	1.71	-0.06	
Ideal SE			1.67		0.85	
5dB	1	-0.94	2.28		1.78	2.41
	2		3.26	3.05	2.65	
	3		3.80	3.78	3.09	
	4		4.05	4.16	3.30	
	5		4.31	4.43	3.50	
	6		4.49	4.63	3.66	
Ideal SE			5.03		4.60	
10dB	1	4.06	6.67		6.37	6.79
	2		7.36	7.31	7.04	
	3		7.75	7.87	7.37	
	4		8.02	8.21	7.55	
	5		8.04	8.30	7.65	
	6		8.11	8.39	7.76	
Ideal SE			8.79		8.54	
15dB	1	9.06	10.87		10.81	10.95
	2		11.18	11.33	11.20	
	3		11.27	11.50	11.32	
	4		11.40	11.64	11.43	
	5		11.45	11.71	11.48	
	6		11.41	11.73	11.50	
Ideal SE			12.19		12.22	
20dB	1	14.06	15.27		15.32	15.31
	2		15.38	15.57	15.52	
	3		15.35	15.63	15.55	
	4		15.36	15.66	15.59	
	5		15.39	15.68	15.61	
	6		15.37	15.68	15.61	
Ideal SE			15.89		15.98	

Table 3.16: Evaluation of recursive speech enhancement approaches: Segmental SNR (Male Speech 2)

Input SNR	Iteration	Input Signal SegSNR	Segmental SNR			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	-7.87	-3.61		-4.48	-3.41
	2		-1.85	-2.48	-3.13	
	3		-0.79	-1.15	-2.37	
	4		-0.05	-0.28	-0.86	
	5		0.44	0.32	-1.46	
	6		0.67	0.68	-1.16	
Ideal SE			1.24		-0.21	
5dB	1	-2.87	0.80		0.17	0.99
	2		2.08	1.76	1.27	
	3		2.75	2.69	1.85	
	4		3.19	3.23	2.20	
	5		3.51	3.57	2.47	
	6		3.80	3.90	2.67	
Ideal SE			4.30		3.56	
10dB	1	2.13	5.06		4.66	5.19
	2		5.87	5.73	5.42	
	3		6.25	6.31	5.78	
	4		6.49	6.65	6.00	
	5		6.67	6.86	6.17	
	6		6.76	6.96	6.28	
Ideal SE			7.65		7.22	
15dB	1	7.13	9.37		9.15	9.47
	2		9.79	9.84	9.64	
	3		9.97	10.15	9.84	
	4		10.08	10.32	9.96	
	5		10.14	10.41	10.04	
	6		10.20	10.47	10.11	
Ideal SE			11.11		10.91	
20dB	1	12.13	13.68		13.61	13.76
	2		13.81	13.98	13.88	
	3		13.86	14.13	14.00	
	4		13.90	14.20	14.07	
	5		13.93	14.24	14.11	
	6		13.92	14.23	14.11	
Ideal SE			14.58		14.54	

Table 3.17: Evaluation of recursive speech enhancement approaches: Segmental SNR (Male Speech 3)

Input SNR	Iteration	Input Signal LLR	Log Likelihood Ratio			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	2.02	1.71		1.80	1.68
	2		1.56	1.59	1.66	
	3		1.62	1.48	1.58	
	4		1.67	1.56	1.52	
	5		1.77	1.54	1.47	
	6		1.79	1.47	1.50	
Ideal SE			1.35		1.33	
5dB	1	1.66	1.32		1.41	1.30
	2		1.22	1.23	1.29	
	3		1.25	1.17	1.21	
	4		1.27	1.19	1.17	
	5		1.33	1.19	1.14	
	6		1.42	1.19	1.13	
Ideal SE			0.999		0.95	
10dB	1	1.29	1.02		1.07	1.00
	2		0.95	0.94	0.98	
	3		1.07	0.92	0.94	
	4		1.11	1.01	0.90	
	5		1.16	1.11	0.90	
	6		1.16	1.16	0.91	
Ideal SE			0.76		0.69	
15dB	1	0.91	0.65		0.70	0.64
	2		0.60	0.59	0.62	
	3		0.71	0.60	0.58	
	4		0.80	0.69	0.56	
	5		0.85	0.74	0.57	
	6		0.86	0.77	0.59	
Ideal SE			0.50		0.44	
20dB	1	0.60	0.42		0.45	0.41
	2		0.41	0.39	0.40	
	3		0.46	0.40	0.39	
	4		0.52	0.46	0.38	
	5		0.57	0.53	0.39	
	6		0.62	0.57	0.40	
Ideal SE			0.35		0.30	

Table 3.18: Evaluation of recursive speech enhancement approaches: Log Likelihood Ratio (Female Speech 1)

Input SNR	Iteration	Input Signal LLR	Log Likelihood Ratio			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	1.61	1.48		1.51	1.47
	2		1.42	1.45	1.46	
	3		1.61	1.44	1.43	
	4		1.66	1.48	1.40	
	5		1.77	1.51	1.40	
	6		1.74	1.59	1.42	
Ideal SE			1.41		1.18	
5dB	1	1.31	1.15		1.18	1.14
	2		1.18	1.10	1.11	
	3		1.33	1.17	1.09	
	4		1.48	1.26	1.09	
	5		1.48	1.41	1.12	
	6		1.43	1.40	1.16	
Ideal SE			1.11		0.90	
10dB	1	1.06	0.87		0.90	0.85
	2		0.86	0.81	0.83	
	3		0.97	0.82	0.80	
	4		1.10	0.92	0.78	
	5		1.17	1.02	0.79	
	6		1.19	1.08	0.81	
Ideal SE			0.77		0.65	
15dB	1	0.78	0.62		0.65	0.60
	2		0.64	0.58	0.59	
	3		0.74	0.61	0.57	
	4		0.86	0.70	0.57	
	5		0.87	0.74	0.56	
	6		0.81	0.76	0.56	
Ideal SE			0.61		0.46	
20dB	1	0.56	0.43		0.45	0.43
	2		0.45	0.42	0.42	
	3		0.50	0.43	0.41	
	4		0.55	0.49	0.40	
	5		0.63	0.54	0.42	
	6		0.63	0.59	0.43	
Ideal SE			0.44		0.32	

Table 3.19: Evaluation of recursive speech enhancement approaches: Log Likelihood Ratio (Female Speech 2)

Input SNR	Iteration	Input Signal LLR	Log Likelihood Ratio			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	2.12	1.80		1.90	1.77
	2		1.51	1.66	1.75	
	3		1.58	1.51	1.64	
	4		1.64	1.46	1.56	
	5		1.71	1.47	1.59	
	6		1.74	1.51	1.67	
Ideal SE			1.32		1.35	
5dB	1	1.72	1.40		1.48	1.37
	2		1.21	1.26	1.33	
	3		1.22	1.15	1.23	
	4		1.30	1.19	1.18	
	5		1.37	1.20	1.19	
	6		1.44	1.21	1.25	
Ideal SE			0.96		1.02	
10dB	1	1.31	1.00		1.06	0.97
	2		0.90	0.91	0.95	
	3		0.89	0.85	0.90	
	4		0.94	0.86	0.86	
	5		1.01	0.88	0.87	
	6		1.08	0.90	0.87	
Ideal SE			0.75		0.70	
15dB	1	0.96	0.70		0.75	0.67
	2		0.66	0.64	0.67	
	3		0.73	0.66	0.64	
	4		0.79	0.71	0.62	
	5		0.84	0.78	0.63	
	6		0.85	0.84	0.65	
Ideal SE			0.50		0.48	
20dB	1	0.64	0.46		0.49	0.44
	2		0.47	0.43	0.44	
	3		0.50	0.46	0.43	
	4		0.57	0.51	0.43	
	5		0.58	0.55	0.43	
	6		0.60	0.56	0.44	
Ideal SE			0.39		0.32	

Table 3.20: Evaluation of recursive speech enhancement approaches: Log Likelihood Ratio (Female Speech 3)

Noise SNR	Iteration	Input Signal LLR	Log Likelihood Ratio			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	2.49	2.09		2.20	2.06
	2		1.93	1.96	2.05	
	3		1.91	1.84	1.94	
	4		1.96	1.79	1.89	
	5		2.06	1.72	1.92	
	6		2.11	1.63	1.98	
Ideal SE			1.57		1.60	
5dB	1	2.01	1.57		1.70	1.55
	2		1.42	1.46	1.54	
	3		1.47	1.35	1.44	
	4		1.46	1.35	1.33	
	5		1.47	1.35	1.38	
	6		1.50	1.37	1.40	
Ideal SE			1.19		1.17	
10dB	1	1.58	1.20		1.29	1.18
	2		1.11	1.11	1.17	
	3		1.13	1.04	1.10	
	4		1.17	1.06	1.06	
	5		1.20	1.12	1.07	
	6		1.22	1.16	1.08	
Ideal SE			0.82		0.85	
15dB	1	1.17	0.85		0.92	0.83
	2		0.79	0.78	0.82	
	3		0.86	0.77	0.77	
	4		0.92	0.80	0.74	
	5		0.95	0.87	0.75	
	6		1.00	0.89	0.75	
Ideal SE			0.62		0.58	
20dB	1	0.81	0.57		0.62	0.56
	2		0.55	0.52	0.54	
	3		0.63	0.52	0.51	
	4		0.69	0.57	0.51	
	5		0.72	0.63	0.50	
	6		0.72	0.65	0.51	
Ideal SE			0.45		0.42	

Table 3.21: Evaluation of recursive speech enhancement approaches: Log Likelihood Ratio (Male Speech 1)

Input SNR	Iteration	Input Signal LLR	Log Likelihood Ratio			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	1.38	1.19		1.23	1.19
	2		1.16	1.12	1.15	
	3		1.22	1.16	1.13	
	4		1.32	1.24	1.09	
	5		1.42	1.20	1.10	
	6		1.44	1.17	1.12	
Ideal SE			0.98		0.90	
5dB	1	1.08	0.86		0.90	0.84
	2		0.81	0.79	0.83	
	3		0.94	0.78	0.78	
	4		0.95	0.85	0.75	
	5		1.05	0.95	0.76	
	6		1.11	1.01	0.76	
Ideal SE			0.69		0.63	
10dB	1	0.78	0.59		0.63	0.58
	2		0.57	0.54	0.55	
	3		0.61	0.55	0.53	
	4		0.75	0.61	0.53	
	5		0.78	0.70	0.53	
	6		0.78	0.77	0.54	
Ideal SE			0.48		0.43	
15dB	1	0.53	0.40		0.41	0.40
	2		0.42	0.38	0.39	
	3		0.50	0.42	0.38	
	4		0.53	0.47	0.38	
	5		0.62	0.51	0.39	
	6		0.60	0.55	0.41	
Ideal SE			0.39		0.31	
20dB	1	0.36	0.30		0.29	0.28
	2		0.31	0.27	0.27	
	3		0.37	0.31	0.27	
	4		0.43	0.35	0.27	
	5		0.45	0.40	0.28	
	6		0.46	0.43	0.29	
Ideal SE			0.29		0.21	

Table 3.22: Evaluation of recursive speech enhancement approaches: Log Likelihood Ratio (Male Speech 2)

Noise SNR	Iteration	Input Signal LLR	Log Likelihood Ratio			
			SE in [118]			SE with Wiener filter
			ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1	1.67	1.52		1.55	1.50
	2		1.51	1.46	1.49	
	3		1.67	1.47	1.45	
	4		1.73	1.59	1.44	
	5		1.83	1.65	1.52	
	6		1.92	1.76	1.63	
Ideal SE			1.25		1.16	
5dB	1	1.35	1.13		1.19	1.12
	2		1.06	1.06	1.10	
	3		1.21	1.08	1.05	
	4		1.23	1.15	1.04	
	5		1.32	1.21	1.06	
	6		1.39	1.31	1.08	
Ideal SE			0.96		0.86	
10dB	1	1.08	0.89		0.91	0.87
	2		0.86	0.83	0.85	
	3		0.95	0.82	0.82	
	4		1.07	0.90	0.82	
	5		1.11	1.02	0.81	
	6		1.10	1.06	0.83	
Ideal SE			0.81		0.65	
15dB	1	0.79	0.63		0.65	0.62
	2		0.63	0.60	0.61	
	3		0.66	0.61	0.59	
	4		0.76	0.67	0.59	
	5		0.78	0.70	0.59	
	6		0.78	0.71	0.59	
Ideal SE			0.54		0.46	
20dB	1	0.55	0.42		0.44	0.42
	2		0.42	0.39	0.40	
	3		0.45	0.39	0.38	
	4		0.54	0.46	0.38	
	5		0.56	0.49	0.39	
	6		0.59	0.51	0.40	
Ideal SE			0.39		0.31	

Table 3.23: Evaluation of recursive speech enhancement approaches: Log Likelihood Ratio (Male Speech 3)

Noise SNR	Iteration	Remaining Energy (%)	PESQ	Segmental SNR	LLR
0dB	0	199.7	1.25	-6.76	2.12
	1	119.6	+0.34	-2.58	1.80
	2	102.6	+0.47	-0.88	1.51
	3	96.5	+0.48	0.15	1.58
	4	93.8	+0.35	0.82	1.64
	5	92.1	+0.16	1.39	1.71
	6	90.9	+0.06	1.68	1.84
Ideal SE		95.3	+0.37	1.95	1.32
5dB	0	131.7	1.60	-1.76	1.72
	1	104.7	+0.39	1.93	1.40
	2	100.0	+0.53	3.29	1.21
	3	98.2	+0.47	3.99	1.22
	4	97.1	+0.27	4.36	1.30
	5	96.3	+0.10	4.67	1.37
	6	95.5	-0.02	4.85	1.44
Ideal SE		97.2	+0.41	5.40	0.96
10dB	0	110.1	1.96	3.24	1.31
	1	100.8	+0.40	6.38	1.00
	2	99.4	+0.54	7.29	0.90
	3	98.6	+0.51	7.75	0.89
	4	98.4	+0.32	8.05	0.94
	5	98.3	+0.17	8.22	1.01
	6	98.0	+0.11	8.32	1.08
Ideal SE		98.4	+0.48	8.96	0.75
15dB	0	103.5	2.33	8.24	0.96
	1	100.0	+0.43	10.63	0.70
	2	99.7	+0.52	11.08	0.66
	3	99.5	+0.55	11.27	0.73
	4	99.4	+0.54	11.36	0.79
	5	99.4	+0.51	11.47	0.84
	6	99.3	+0.49	11.51	0.85
Ideal SE		99.4	+0.63	12.22	0.70
20dB	0	101.1	2.70	13.24	0.64
	1	99.9	+0.39	15.01	0.46
	2	99.8	+0.49	15.26	0.47
	3	99.7	+0.51	15.33	0.50
	4	99.7	+0.51	15.33	0.57
	5	99.7	+0.50	15.35	0.58
	6	99.7	+0.48	15.35	0.60
Ideal SE		99.6	+0.58	15.76	0.39

Table 3.24: Evaluation of recursive speech enhancement after each iteration with ML approach (Female Speech 3). Iteration 0 means the input noisy signal, PESQ scores for interation 1 to 6 are the PESQ score improvement comparing to input noisy signal.

Noise SNR	Iteration	Remaining Energy (%)	PESQ	Segmental SNR	LLR
0dB	0	199.7	1.25	-6.76	2.12
	1	130.5	+0.26	-3.41	1.90
	2	114.9	+0.41	-2.09	1.75
	3	108.5	+0.52	-1.32	1.64
	4	105.8	+0.59	-0.83	1.56
	5	103.1	+0.53	-0.46	1.59
	6	101.5	+0.33	-0.17	1.67
Ideal SE		101.3	+0.64	0.80	1.35
5dB	0	131.7	1.60	-1.76	1.72
	1	108.2	+0.30	1.28	1.48
	2	103.8	+0.46	2.41	1.33
	3	102.2	+0.54	2.97	1.23
	4	100.9	+0.57	3.29	1.18
	5	100.0	+0.51	3.55	1.19
	6	99.2	+0.25	3.79	1.25
Ideal SE		99.3	+0.70	4.68	1.02
10dB	0	110.1	1.96	3.24	1.31
	1	101.9	+0.29	5.90	1.06
	2	100.6	+0.46	6.73	0.95
	3	99.9	+0.54	7.10	0.90
	4	99.6	+0.54	7.33	0.86
	5	99.5	+0.49	7.51	0.85
	6	99.2	+0.39	7.59	0.85
Ideal SE		99.1	+0.69	8.53	0.70
15dB	0	103.5	2.33	8.24	0.96
	1	100.5	+0.33	10.39	0.75
	2	100.0	+0.44	10.93	0.67
	3	99.9	+0.51	11.15	0.64
	4	99.9	+0.54	11.25	0.62
	5	99.8	+0.54	11.34	0.63
	6	99.7	+0.54	11.39	0.65
Ideal SE		99.7	+0.70	12.09	0.48
20dB	0	101.1	2.70	13.24	0.64
	1	99.9	+0.30	14.90	0.49
	2	99.9	+0.38	15.23	0.44
	3	99.9	+0.43	15.39	0.43
	4	99.8	+0.44	15.42	0.43
	5	99.8	+0.44	15.47	0.43
	6	99.8	+0.43	15.49	0.44
Ideal SE		99.7	+0.62	15.74	0.32

Table 3.25: Evaluation of recursive speech enhancement after each iteration with MMSE approach (Female Speech 3). Iteration 0 means the input noisy signal, PESQ scores for interation 1 to 6 are the PESQ score improvement comparing to input noisy signal.

Based on the results in Tables 3.6 to 3.23, comparing the performance using PESQ, Segmental SNR and Log Likelihood Ratio, we can see that:

• If we use ML results as the reference for the next ML estimation, the PESQ score will reach its maximum value after the second or third iteration. This is due to the fact that with each iteration, the remaining energy of enhanced speech is reduced as shown in Table 3.24. This leads to smaller estimate of a compared to the previous iteration. The new estimate of a tends to remove more noise, which is confirmed by the segmental SNR improvement. After the second or third iteration, the enhanced speech has significantly lower energy than the original clean speech. This means that the under-estimation of a results in removing more noise than necessary and hence higher distortion (larger LLR values). Even though the noise is further reduced after each iteration, the increased distortion will overshadow the effect of noise removal after the second or third iteration. The results after the second iteration usually outperform the Wiener filter approach.

• When the MMSE estimation approach is used to follow an ML enhanced signal, there is slightly lower PESQ improvement than for the ML on ML approach. Similarly, this approach will reach the maximum performance after the 3rd iteration. This is also due to the compromise of noise removal and distortion. In most cases, especially the low SNR cases, this approach will outperform the ML on ML approach.

• When the MMSE estimation follows an MMSE enhanced signal, the PESQ score improves after each of the first four iterations. The scores usually outperform the Wiener filter approach after the second iteration. Overall the MMSE estimation has the best possible performance after around 4 iterations. From the Table 3.25, the remaining energy of the enhanced signal decreases significantly in the first three iterations. We already concluded

that smaller a leads to more aggressive noise removal. Thus for the low SNR case, even though the overall energy may still be higher than 100%, the energy at lower SNR frames are lower than others. For these frames, the a is considerably under-estimated. This effect leads to increasing distortion and lower PESQ scores after the 4th iteration.

• Comparing the three approaches, we can conclude that the ML approach removes more noise than the MMSE approach. This leads to higher distortion for some low SNR frames. As shown in the previous section, the overall energy at each frame will be reduced at each iteration. Hence the estimation of a tends to decrease and the overall energy will be further reduced at the next iteration. Comparing to MMSE approach, ML leads to more under-estimation and higher chance of distortion. Thus in the first 2 iterations, ML provides better quality due to its quick removal of the noise, but MMSE preserves the overall speech better. MMSE does not remove the noise as fast as the ML approach, but it keeps the distortion lower. After the first 2 iterations, MMSE outperforms ML in the overall quality and distortion measures. On the other hand, ML generally provides higher segmental SNRs than MMSE.

• Overall, the MMSE on MMSE approach achieves the best quality. It reaches the best performance after the 4th iteration. The MMSE on ML results are comparable with the MMSE on MMSE approach. This means ML approach can be used first for fast removal of noise, and MMSE approach can be used later to minimize error and achieve better performance. It will increase the computational complexity compared to the other approaches. Alternative approaches for the proposed algorithm can be: ML estimation with 2 iteration to achieve low cost and good performance, MMSE on the ML estimation with 3 iterations overall (MMSE on ML on ML) to achieve higher performance with higher cost.

In Tables 3.26 to 3.31, the composite quality measures are used to evaluate the performance of the speech enhancement. The composite quality measures used to evaluate the performance are:

$$M_{overall} = 1.594 - 0.512 \cdot LLR + 0.805 \cdot PESQ \tag{3.19}$$

$$M_{back} = 1.634 + 0.063 \cdot segSNR + 0.478 \cdot PESQ \tag{3.20}$$

$$M_{sig} = 3.093 - 1.029 \cdot LLR + 0.603 \cdot PESQ \tag{3.21}$$

where a higher value implies better quality. $M_{overall}$, M_{back} and M_{sig} are the evaluations of the overall quality, the background noise and the speech signal, respectively.

We choose to compare the results of three recursive approaches: ML on ML after 2nd iteration, MMSE on ML after 3rd iteration and MMSE on MMSE after 4th iteration. The results confirms that our proposed recursive approach outperforms the Wiener filter approach. The MMSE on ML and MMSE on MMSE approaches achieve similar improvements overall. MMSE on ML leads to better background noise reduction and MMSE on MMSE leads to better signal preservation. All proposed approaches significantly improve over the Wiener filter approach. Next, we will evaluate the proposed algorithms with non-white noises.

Female Speech 1

Noise SNR	Input Signal $M_{overall}$	$M_{overall}$			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.65	2.12	2.21	2.22	2.00
5dB	2.07	2.63	2.69	2.70	2.50
10dB	2.51	3.03	3.07	3.09	2.93
15dB	3.00	3.53	3.55	3.56	3.43
20dB	3.43	3.90	3.91	3.92	3.81

Female Speech 2

Noise SNR	Input Signal $M_{overall}$	$M_{overall}$			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	2.07	2.37	2.42	2.44	2.31
5dB	2.41	2.74	2.78	2.82	2.74
10dB	2.81	3.22	3.27	3.28	3.16
15dB	3.15	3.55	3.59	3.58	3.50
20dB	3.50	3.90	3.90	3.90	3.80

Female Speech 3

Noise SNR	Input Signal $M_{overall}$	$M_{overall}$			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.53	2.21	2.26	2.28	1.98
5dB	2.01	2.69	2.75	2.75	2.51
10dB	2.50	3.15	3.18	3.17	3.01
15dB	2.99	3.55	3.56	3.59	3.47
20dB	3.42	3.92	3.90	3.90	3.85

Table 3.26: Evaluation of recursive speech enhancement approaches: $M_{overall}$ (Female Speech)

Male Speech 1

Noise SNR	Input Signal $M_{overall}$	$M_{overall}$			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.44	1.93	2.03	2.07	1.77
5dB	1.82	2.48	2.55	2.60	2.28
10dB	2.27	2.92	2.98	2.98	2.78
15dB	2.72	3.31	3.35	3.37	3.21
20dB	3.17	3.72	3.75	3.76	3.63

Male Speech 2

Noise SNR	Input Signal $M_{overall}$	$M_{overall}$			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	2.19	2.59	2.64	2.73	2.49
5dB	2.62	3.05	3.13	3.11	2.96
10dB	3.02	3.48	3.51	3.53	3.44
15dB	3.48	3.85	3.87	3.89	3.85
20dB	3.92	4.15	4.14	4.16	4.10

Male Speech 3

Noise SNR	Input Signal $M_{overall}$	$M_{overall}$			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.83	2.27	2.37	2.39	2.19
5dB	2.25	2.82	2.88	2.91	2.70
10dB	2.71	3.21	3.26	3.25	3.15
15dB	3.18	3.68	3.70	3.70	3.60
20dB	3.60	4.00	4.01	4.01	3.94

Table 3.27: Evaluation of recursive speech enhancement approaches: $M_{overall}$ (Male Speech)

Female Speech 1

Noise SNR	Input Signal M_{back}	M_{back}			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.87	2.51	2.58	2.47	2.28
5dB	2.32	2.86	2.92	2.89	2.73
10dB	2.77	3.25	3.29	3.28	3.16
15dB	3.27	3.68	3.71	3.70	3.61
20dB	3.76	4.09	4.11	4.11	4.03

Female Speech 2

Noise SNR	Input Signal M_{back}	M_{back}			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.66	2.21	2.29	2.24	2.06
5dB	2.10	2,61	2.68	2.63	2.50
10dB	2.56	3.05	3.09	3.06	2.94
15dB	3.00	3.42	3.47	3.43	3.34
20dB	3.46	3.82	3.84	3.81	3.73

Female Speech 3

Noise SNR	Input Signal M_{back}	M_{back}			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.81	2.40	2.48	2.46	2.25
5dB	2.29	2.86	2.92	2.88	2.73
10dB	2.78	3.29	3.32	3.29	3.18
15dB	3.27	3.69	3.72	3.71	3.63
20dB	3.76	4.12	4.12	4.10	4.05

Table 3.28: Evaluation of recursive speech enhancement approaches: M_{back} (Female Speech)

Male Speech 1

Noise SNR	Input Signal M_{back}	M_{back} SE in [118] ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	SE with Wiener filter
0dB	1.90	2.43	2.52	2.51	2.27
5dB	2.30	2.88	2.95	2.92	2.71
10dB	2.74	3.29	3.34	3.32	3.16
15dB	3.21	3.67	3.72	3.70	3.59
20dB	3.69	4.09	4.12	4.11	4.02

Male Speech 2

Noise SNR	Input Signal M_{back}	M_{back} SE in [118] ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	SE with Wiener filter
0dB	2.04	2.54	2.61	2.61	2.41
5dB	2.51	2.95	3.02	2.97	2.85
10dB	2.98	3.39	3.43	3.42	3.33
15dB	3.49	3.81	3.84	3.83	3.79
20dB	4.01	4.22	4.22	4.22	4.17

Male Speech 3

Noise SNR	Input Signal M_{back}	M_{back} SE in [118] ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	SE with Wiener filter
0dB	1.78	2.38	2.47	2.49	2.23
5dB	2.27	2.82	2.90	2.87	2.69
10dB	2.75	3.23	3.27	3.25	3.15
15dB	3.26	3.68	3.71	3.70	3.62
20dB	3.75	4.06	4.08	4.07	4.02

Table 3.29: Evaluation of recursive speech enhancement approaches: M_{back} (Male Speech)

Female Speech 1

Noise SNR	Input Signal M_{sig}	M_{sig}			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.83	2.44	2.55	2.58	2.31
5dB	2.38	3.09	3.16	3.17	2.93
10dB	2.98	3.56	3.61	3.63	3.45
15dB	3.57	4.16	4.17	4.21	4.06
20dB	4.07	4.55	4.57	4.59	4.49

Female Speech 2

Noise SNR	Input Signal M_{sig}	M_{sig}			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	2.42	2.76	2.78	2.82	2.68
5dB	2.83	3.19	3.23	3.30	3.22
10dB	3.34	3.76	3.82	3.85	3.72
15dB	3.74	4.15	4.20	4.21	4.13
20dB	4.15	4.53	4.54	4.56	4.47

Female Speech 3

Noise SNR	Input Signal M_{sig}	M_{sig}			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.70	2.58	2.62	2.60	2.24
5dB	2.30	3.13	3.22	3.19	2.90
10dB	2.93	3.67	3.73	3.73	3.52
15dB	3.54	4.13	4.14	4.19	4.07
20dB	4.03	4.53	4.53	4.54	4.50

Table 3.30: Evaluation of recursive speech enhancement approaches: M_{sig} (Female Speech)

Male Speech 1

Noise SNR	Input Signal M_{sig}	M_{sig} SE in [118] ML after ML(s)	M_{sig} SE in [118] MMSE after ML(s)	M_{sig} SE in [118] MMSE after MMSE(s)	M_{sig} SE with Wiener filter
0dB	1.37	2.10	2.23	2.23	1.90
5dB	1.95	2.83	2.94	2.98	2.61
10dB	2.59	3.37	3.45	3.45	3.21
15dB	3.16	3.87	3.91	3.95	3.77
20dB	3.74	4.33	4.37	4.38	4.35

Male Speech 2

Noise SNR	Input Signal M_{sig}	M_{sig} SE in [118] ML after ML(s)	M_{sig} SE in [118] MMSE after ML(s)	M_{sig} SE in [118] MMSE after MMSE(s)	M_{sig} SE with Wiener filter
0dB	2.65	3.09	3.13	3.24	3.00
5dB	3.17	3.66	3.74	3.74	3.57
10dB	3.63	4.14	4.17	4.20	4.10
15dB	4.16	4.51	4.53	4.57	4.52
20dB	4.61	4.81	4.80	4.84	4.79

Male Speech 3

Noise SNR	Input Signal M_{sig}	M_{sig} SE in [118] ML after ML(s)	M_{sig} SE in [118] MMSE after ML(s)	M_{sig} SE in [118] MMSE after MMSE(s)	M_{sig} SE with Wiener filter
0dB	2.19	2.62	2.73	2.76	2.57
5dB	2.69	3.33	3.36	3.40	3.19
10dB	3.24	3.75	3.81	3.81	3.69
15dB	3.78	4.25	4.27	4.29	4.20
20dB	4.24	4.62	4.65	4.66	4.58

Table 3.31: Evaluation of recursive speech enhancement approaches: M_{sig} (Male Speech)

Tables 3.32 to 3.35 shows the performances of the recursive enhancement algorithms based on Laplacian-Gaussian model under non-white noises. Two types of noises are used, the F16 noise and the babble noise. Both noise samples are chosen from the NOISEX-92 database (NOISE-ROM-0 signal 011 and 020) [16] . The NOISEX-92 database was originally sampled at 19.98 KHz. It is re-sampled with Matlab to 8 KHz to fit with our speech samples.

For the F16 noise, the improvement is significant for PESQ and segmental SNR. For the LLR measure, the MMSE after MMSE approach achieves better results than the Wiener filter approach, while the MMSE after ML approach is comparable with the Wiener filter approach.

For the Babble noise, the improvement is not as significant comparing to the noisy speech for the PESQ measure. The segmental SNR results show that the noise reduction is still significant, but the LLR results show slightly higher distortion. This is due to the fact that additive babble noise has frequency characteristics similar to the speech itself. This makes it more difficult to separate them one from the other. Comparing to the Wiener filter approach, our proposed recursive approaches still outperform for PESQ and segmental SNR measures. The MMSE after MMSE approach can still lead to lower LLR than the Wiener filter approach, while the other two approaches (ML after ML and MMSE after ML) are comparable or slightly worse measured by the LLR.

Overall, preliminary results shows that the proposed recursive algorithms outperform the Wiener filter method.

Noise SNR	Input Signal PESQ	PESQ improvement			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.47	+0.07	+0.11	+0.14	+0.10
5dB	1.90	+0.13	+0.14	+0.15	+0.14
10dB	2.30	+0.14	+0.17	+0.17	+0.14
15dB	2.70	+0.14	+0.15	+0.16	+0.14
20dB	3.06	+0.18	+0.19	+0.20	+0.15

Noise SNR	Input Seg-mental SNR	Segmental SNR			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	-6.58	-2.53	-2.11	-2.50	-3.09
5dB	-1.58	1.37	1.76	1.50	1.13
10dB	3.42	5.25	5.66	5.52	5.32
15dB	8.42	9.48	9.83	9.75	9.63
20dB	13.42	13.98	14.24	14.21	14.11

Noise SNR	Input Signal LLR	Log Likelihood Ratio			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	0.95	1.05	1.05	0.99	0.95
5dB	0.74	0.76	0.74	0.70	0.70
10dB	0.53	0.53	0.52	0.51	0.52
15dB	0.36	0.37	0.37	0.35	0.36
20dB	0.23	0.26	0.24	0.23	0.25

Table 3.32: Comaprison of objective quality measures for recursive speech enhancement approaches (Female Speech 3, Babble noise)

Noise SNR	Input Signal PESQ	PESQ improvement			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.66	+0.14	+0.16	+0.16	+0.14
5dB	2.02	+0.17	+0.20	+0.20	+0.15
10dB	2.37	+0.20	+0.23	+0.23	+0.18
15dB	2.74	+0.26	+0.29	+0.30	+0.22
20dB	3.13	+0.22	+0.23	+0.22	+0.17

Noise SNR	Input Seg-mental SNR	Segmental SNR			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	-7.84	-3.50	-3.01	-3.47	-4.39
5dB	-2.84	0.51	1.00	0.66	-0.07
10dB	2.16	4.60	5.04	4.82	4.29
15dB	7.16	8.95	9.27	9.13	8.76
20dB	12.16	13.44	13.69	13.57	13.32

Noise SNR	Input Signal LLR	Log Likelihood Ratio			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.07	1.10	1.09	1.05	1.05
5dB	0.82	0.80	0.78	0.76	0.76
10dB	0.60	0.56	0.53	0.51	0.54
15dB	0.42	0.37	0.34	0.34	0.36
20dB	0.28	0.24	0.22	0.22	0.24

Table 3.33: Comaprison of objective quality measures for recursive speech enhancement approaches (Male Speech 3, Babble noise)

Noise SNR	Input Signal PESQ	PESQ improvement			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.45	+0.22	+0.26	+0.30	+0.17
5dB	1.87	+0.30	+0.33	+0.35	+0.26
10dB	2.28	+0.29	+0.31	+0.34	+0.24
15dB	2.67	+0.32	+0.32	+0.35	+0.25
20dB	3.03	+0.30	+0.33	+0.34	+0.21

Noise SNR	Input Seg-mental SNR	Segmental SNR			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	-5.94	-0.42	0.08	-0.65	-1.83
5dB	-0.94	3.27	3.70	3.18	2.42
10dB	4.06	6.98	7.33	7.01	6.56
15dB	9.06	10.92	11.24	11.05	10.78
20dB	14.06	15.16	15.43	15.31	15.11

Noise SNR	Input Signal LLR	Log Likelihood Ratio			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.24	1.10	1.09	1.03	1.08
5dB	0.97	0.82	0.81	0.76	0.80
10dB	0.70	0.58	0.58	0.54	0.57
15dB	0.48	0.41	0.40	0.38	0.39
20dB	0.30	0.27	0.26	0.25	0.25

Table 3.34: Comaprison of objective quality measures for recursive speech enhancement approaches (Female Speech 3, F16 noise)

Noise SNR	Input Signal PESQ	PESQ improvement			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.66	+0.30	+0.33	+0.36	+0.23
5dB	1.99	+0.31	+0.32	+0.35	+0.23
10dB	2.37	+0.32	+0.34	+0.34	+0.25
15dB	2.75	+0.32	+0.33	+0.33	+0.24
20dB	3.12	+0.30	+0.30	+0.31	+0.22

Noise SNR	Input Seg-mental SNR	Segmental SNR			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	-7.10	-1.58	-0.99	-1.51	-2.86
5dB	-2.10	2.19	2.67	2.35	1.36
10dB	2.90	5.98	6.39	6.22	5.55
15dB	7.90	9.95	10.30	10.23	9.78
20dB	12.90	14.22	14.50	14.47	14.18

Noise SNR	Input Signal LLR	Log Likelihood Ratio			
		SE in [118]			SE with Wiener filter
		ML after ML(s)	MMSE after ML(s)	MMSE after MMSE(s)	
0dB	1.14	1.16	1.13	1.09	1.12
5dB	0.89	0.87	0.83	0.80	0.83
10dB	0.67	0.61	0.59	0.57	0.60
15dB	0.49	0.40	0.40	0.39	0.41
20dB	0.33	0.27	0.26	0.26	0.27

Table 3.35: Comaprison of objective quality measures for recursive speech enhancement approaches (Male Speech 3, F16 noise)

3.6 Chapter Summary

In this chapter, a speech enhancement algorithm based on Laplacian-Gaussian model (for the speech and noise, respectively) was reviewed. Unlike other classical speech enhancement algorithms, the speech signal was assumed to follow Laplacian distribution instead of Gaussian. The speech enhancement is achieved by estimating the Laplacian components from a Gaussian plus Laplacian mixture. The speech Laplacian factor and the variance of the noise variance were estimated adaptively using the noisy signal and the noise only signal. The speech enhancement has been designed based on the MMSE or ML approach.

The estimation of the speech Laplacian factor is very critical for the performance of the algorithm. This factor was estimated from the noisy speech in [118]. Noting that the speech Laplacian factor estimation can be improved by using the enhanced speech signal which has less noise, instead of the noisy speech itself, a recursive speech enhancement algorithm was proposed. The enhanced speech signal was used as a reference signal to estimate the Laplacian factor, recursively. The updated Laplacian factors is then used to re-enhance the original noisy speech signal. This provides better estimation and leads to better speech enhancement in the next iteration. The simulation results proved that this recursive approach can lead to further improved performance of enhancing the speech.

The performance of the algorithm is assessed using PESQ, segmental SNR and LLR quality measures. With each iteration, more noise is removed from the noisy signal. This effect results in an increasing segmental SNR. At the same time, it begins to distort the speech signal itself (shown by the increase in the distortion measure: LLR), usually after the 2nd iteration for ML and the 4th or 5th iteration for MMSE. The overall performance (indicated by PESQ measure) compromises between the previous two measures and achieves the best quality after the second iteration for ML estimation and after the 4th

iteration for MMSE. Another alternative is to apply MMSE after 2 iterations of ML estimation. The algorithms tend to perform worse because the loss of energy at each iterations eventually led to under-estimation of a and higher distortion of the speech.

Overall, the PESQ score and the informal listening tests both confirm that the quality of the enhanced speech can be improved with the proposed recursive approach clearly outperforming the Wiener filter approach.

Chapter 4

Speech Enhancement based on Adaptive-α Subtraction Model

4.1 Introduction

In most classical speech enhancement algorithms, the focus is on removal of the additive noise [32] [34] [80]. Spectral subtraction has been used as an efficient method for speech enhancement [26] [80]. This approach estimates the power spectral density of the clean speech signal by subtracting the power spectral density (PSD) of the noise estimate from the PSD of the noisy signal. It is also known that the absolute value of the spectral component can be used instead of the power for subtraction [25] [26] [80]. The envelope of the spectrum [106] has been used as a substitute for PSD in the subtraction. The main drawback of the spectral subtraction methods is the high residual noise level in the enhanced speech, especially when the SNR is low.

In [81] [99], the spectral subtraction approach is generalized by estimating the short

time power spectrum $\left|\hat{X}\right|^2$ as $\left[|Y|^{2\alpha} - |N|^{2\alpha}\right]^{1/\alpha}$. The value of α is chosen to control the amount of subtraction of the algorithm. The value of α is generally selected as 1 for power spectral subtraction and 0.5 for magnitude spectral subtraction [26] [99]. However, the best value of α depends on the SNR, the type of the signal and the listeners. In this Chapter, we focus on the relationship between α and the SNR.

In [10], the relation between the specific loudness and the excitation (as such α) was studied. It was shown that with higher SNR, α tends to increase, suggesting that α should be a function of the SNR. In this chapter, different α values will be tested in the loudness model used for speech enhancement. This is the basis of the Adaptive-α Loudness subtraction approach proposed here.

As the objective quality of speech is highly correlated with the loudness of speech, the lower α value in the range of 0.23 to 0.3, which is suggested by the commonly used loudness measure [82] [10], will be specially considered. We expect these lower α values can be beneficial for spectral subtraction type speech enhancement.

In this Chapter, we will investigate the effect of the various choices of α values. First, results for a fixed α will be examined and then an adaptive-α subtraction algorithm will be proposed for speech enhancement.

4.2 Fixed-α Subtraction Speech Enhancement Algorithm

4.2.1 Fixed-α Subtraction

In this section, the quality of speech enhancement through subtraction with different fixed values of α will be tested. X, Y and N are the magnitude values of the FFT for clean speech, noisy speech and noise only signals, respectively. For each component Y and N, a

clean speech estimate $\hat{X}$ will be calculated as

$$\hat{X} = \left((Y)^{2\alpha} - (N)^{2\alpha}\right)^{1/2\alpha}, \tag{4.1}$$

where the value of α is fixed for all the frames within one test. N is estimated using the noise only intervals within the noisy speech signal. A set of values for $\alpha = 1$, 0.5, 0.33, 0.27 and 0.2 will be tested. The first two values of 1 and 0.5 correspond to the power and magnitude spectral subtraction, respectively.

4.2.2 Simulation Results

In this section, we simulate the classical spectral subtraction, fixed-α subtraction and Wiener filter approaches of speech enhancement.

All the simulations are done in the FFT domain with the same noise estimation process. The speech is sampled at 8 kHz and processed on a frame-by-frame basis. The frame length used here is 128 samples, which corresponds to approximately 16 ms of speech. The speech can be assumed to be stationary for such a frame length. The overlap between successive frames is 64 samples.

In order to understand the performance of speech enhancement algorithms, independent of the performance of noise estimation, a noise reference signal is used instead of estimating the noise from the noisy speech itself. A noise signal that has the same energy level and frequency response as the original noise in the noisy speech is used. The noise is transformed to the FFT domain on the same frame-by-frame basis. The noise variance λ_n is derived with a low-pass filter:

$$\widehat{\lambda_n}(m) = \beta_n \widehat{\lambda_n}(m-1) + (1-\beta_n)|N(m)|^2. \tag{4.2}$$

β_N is chosen to provide a time constant of 0.5 s, assuming slower variation with time for the noise than that of the speech.

The variance of the noisy speech component (used in Wiener filter approach) is also estimated with a low-pass filter.

$$\widehat{\lambda}_y(m) = \beta_n \widehat{\lambda}_y(m-1) + (1-\beta_y)|Y(m)|^2. \tag{4.3}$$

The time constant here is chosen to be 10 ms since speech can be assumed to be stationary within 10-20 ms [8].

The test samples are clean male or female speech corrupted with different levels of white noise. Based on the range of values used in previous works, we choose $\alpha = 1$, 0.5, 0.33, 0.27 and 0.2, respectively.

Three measures are used for the evaluation: ITU p.862 (PESQ), Segmental SNR and Log Likelihood Ratio (LLR) [61]. The details of these measures were discussed in Section 2.4. PESQ is the measure for overall quality, Segmental SNR for noise reduction and LLR for distortion. Results are shown in the following tables.

Tables 4.1 to 4.2 show the PESQ score improvements of the enhanced speech using (4.1). The results show a consistent pattern for different speech files. When the input SNR is low, smaller α values tend to perform better and vice versa. When the SNR is high, $\alpha = 0.5$ leads to the best PESQ improvements. The results suggest that a single α value is not suitable for the full range of input SNRs. To achieve the best performance of the speech enhancement algorithm based on the PESQ assessment, different α's need to be used for different input SNRs.

Tables 4.3 and 4.4 show the segmental SNRs of the enhanced speech for different α values. Again, when the input SNR is low, the three smaller α values provide higher segmental SNRs. When the input SNR is high, the segmental SNR results are better for α

equal to 0.5 and 1.

When the input SNR is low, the larger α values can still improve the quality of speech but the noise removal is not sufficient. The smaller α values removes more noise but the segmental SNR decreases for $\alpha = 0.20$ because the noise removal is overdone. On the other hand, when the input SNR is high, more noise removal leads to deterioration of the speech quality by removing a portion of the clean speech signal. Under such circumstances, larger α values are more suitable because they limit the noise to be removed.

In Tables 4.5 and 4.6, the LLR is used to measure the quality of the enhanced speech. Results show that the performance measured by LLR is consistent with that measured by PESQ or segmental SNR. When the input SNR is low, smaller α values result in better LLR. For some of the speech samples, the impact of changing α is not significant. When the input SNR is high (20 dB), α=0.5 shows the best performance overall.

When the SNR is 0 dB, all tested values reduce the LLR, which means lower distortion. The α values of 0.2, 0.27 and 0.33 lead to similar distortion. When the input SNR increases, the best choice of α value also increases. At 10 dB SNR, $\alpha = 0.33$ performs best while α=0.5 is more suitable for SNRs above 10 dB. When the input SNR is above 10 dB, the smaller α values of 0.2, 0.27 and 0.33 may even lead to higher distortion than the noisy speech.

From Tables 4.1 to 4.6, we can see that the fixed-α subtraction method with smaller α values outperforms the spectral subtraction methods in the low input SNR cases using all three performance measures. On the other hand, higher SNR requires larger α values. Among different chosen α, there is always at least one α that can provide clear improvement over the traditional spectral subtraction (where $\alpha = 1$).

The three objective quality measures used, PESQ, segmental SNR, and LLR all provide

Female Speech 1

Input SNR (dB)	Input Signal PESQ	PESQ score improvements Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	1.36	+0.14	+0.24	+0.27	+0.28	+0.37
5	1.64	+0.17	+0.31	+0.38	+0.40	+0.42
10	1.95	+0.20	+0.36	+0.44	+0.44	+0.38
15	2.30	+0.20	+0.36	+0.37	+0.32	+0.19
20	2.69	+0.20	+0.36	+0.31	+0.21	-0.02

Female Speech 2

Input SNR (dB)	Input Signal PESQ	PESQ score improvements Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	1.57	+0.08	+0.15	+0.16	+0.17	+0.17
5	1.85	+0.11	+0.19	+0.20	+0.19	+0.16
10	2.17	+0.13	+0.21	+0.21	+0.17	+0.02
15	2.47	+0.15	+0.24	+0.21	+0.13	-0.08
20	2.75	+0.15	+0.25	+0.23	+0.16	-0.13

Female Speech 3

Input SNR (dB)	Input Signal PESQ	PESQ score improvements Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	1.22	+0.18	+0.35	+0.46	+0.51	+0.55
5	1.61	+0.22	+0.39	+0.48	+0.49	+0.51
10	1.96	+0.21	+0.38	+0.45	+0.44	+0.32
15	2.34	+0.22	+0.40	+0.44	+0.36	+0.14
20	2.71	+0.19	+0.35	+0.33	+0.22	-0.06

Table 4.1: Comparison of PESQ improvements for subtraction model with different α (Female speech)

Male Speech 1

Input SNR (dB)	Input Signal PESQ	PESQ score improvements Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	1.39	+0.04	+0.06	+0.09	+0.11	+0.15
5	1.59	+0.13	+0.23	+0.30	+0.33	+0.35
10	1.82	+0.19	+0.33	+0.38	+0.38	+0.30
15	2.14	+0.19	+0.32	+0.33	+0.28	+0.08
20	2.51	+0.17	+0.27	+0.20	+0.10	-0.16

Male Speech 2

Input SNR (dB)	Input Signal PESQ	PESQ score improvements Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	1.58	+0.13	+0.22	+0.26	+0.27	+0.29
5	1.93	+0.16	+0.31	+0.38	+0.38	+0.37
10	2.31	+0.21	+0.40	+0.44	+0.41	+0.28
15	2.70	+0.25	+0.37	+0.31	+0.23	+0.02
20	3.18	+0.15	+0.25	+0.12	-0.03	-0.30

Male Speech 3

Input SNR (dB)	Input Signal PESQ	PESQ score improvements Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	1.36	+0.17	+0.32	+0.40	+0.44	+0.52
5	1.69	+0.23	+0.41	+0.55	+0.59	+0.60
10	2.06	+0.24	+0.43	+0.49	+0.51	+0.39
15	2.43	+0.26	+0.43	+0.45	+0.40	+0.27
20	2.83	+0.21	+0.37	+0.37	+0.26	-0.01

Table 4.2: Comparison of PESQ improvements for subtraction model with different α (Male speech)

Female Speech 1

Input SNR (dB)	Input Signal SegSNR	Segmental SNR Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	-6.56	-3.20	0.34	1.51	1.34	0.69
5	-1.56	1.42	4.04	3.62	2.68	1.27
10	3.44	5.81	7.53	5.65	4.05	1.95
15	8.44	10.24	11.03	7.97	5.80	2.94
20	13.44	14.61	14.41	10.27	7.57	4.02

Female Speech 2

Input SNR (dB)	Input Signal SegSNR	Segmental SNR Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	-11.78	-8.49	-4.67	-2.21	-1.28	-0.25
5	-6.78	-3.67	-0.53	0.59	0.68	0.62
10	-1.78	0.91	3.21	2.93	2.36	1.42
15	3.22	5.25	6.76	5.31	4.07	2.23
20	8.22	9.78	10.37	7.56	5.67	3.04

Female Speech 3

Input SNR (dB)	Input Signal SegSNR	Segmental SNR Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	-6.76	-3.57	-0.28	0.77	0.71	0.43
5	-1.76	1.12	3.59	3.09	2.28	1.16
10	3.24	5.55	7.08	5.32	3.89	1.95
15	8.24	10.02	10.77	7.78	5.70	2.93
20	13.24	14.42	14.44	10.36	7.66	4.07

Table 4.3: Comparison of Segmental SNRs for subtraction model with different α (Female speech)

Male Speech 1

Input SNR (dB)	Input Signal SegSNR	Segmental SNR Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	-6.30	-2.91	0.59	1.55	1.37	0.76
5	-1.30	1.70	4.35	3.82	2.86	1.37
10	3.70	6.21	8.03	6.12	4.46	2.17
15	8.70	10.59	11.64	8.58	6.29	3.22
20	13.70	15.02	15.18	10.93	8.08	4.28

Male Speech 2

Input SNR (dB)	Input Signal SegSNR	Segmental SNR Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	-5.95	-3.03	-0.01	0.95	0.92	0.62
5	-0.95	1.57	3.74	3.30	2.52	1.31
10	4.05	6.09	7.41	5.68	4.25	2.22
15	9.05	10.48	10.90	7.89	5.84	3.04
20	14.05	14.93	14.51	10.34	7.70	4.15

Male Speech 3

Input SNR (dB)	Input Signal SegSNR	Segmental SNR Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	-7.87	-4.60	-1.07	0.41	0.59	0.47
5	-2.87	-0.01	2.67	2.71	2.13	1.15
10	2.13	4.41	6.13	4.82	3.60	1.88
15	7.13	8.77	9.51	7.00	5.19	2.74
20	12.13	13.23	13.18	9.45	7.03	3.78

Table 4.4: Comparison of Segmental SNRs for subtraction model with different α (Male speech)

Female Speech 1

Input SNR (dB)	Input Signal LLR	Log Likelihood Ratio (LLR) Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	2.03	1.82	1.68	1.62	1.61	1.61
5	1.67	1.41	1.27	1.21	1.21	1.24
10	1.26	1.06	0.92	0.89	0.91	0.98
15	0.89	0.71	0.61	0.63	0.68	0.79
20	0.60	0.47	0.42	0.47	0.53	0.67

Female Speech 2

Input SNR (dB)	Input Signal LLR	Log Likelihood Ratio (LLR) Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	1.59	1.52	1.44	1.41	1.42	1.46
5	1.33	1.23	1.15	1.15	1.18	1.23
10	1.05	0.94	0.89	0.90	0.93	0.98
15	0.80	0.70	0.68	0.74	0.80	0.88
20	0.55	0.49	0.47	0.53	0.59	0.71

Female Speech 3

Input SNR (dB)	Input Signal LLR	Log Likelihood Ratio (LLR) Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	2.12	1.87	1.70	1.61	1.57	1.53
5	1.72	1.46	1.27	1.18	1.17	1.19
10	1.31	1.12	1.00	0.98	1.00	1.05
15	0.96	0.78	0.68	0.70	0.76	0.86
20	0.64	0.55	0.49	0.55	0.61	0.71

Table 4.5: Comparison of Log Likelihood Ratios for subtraction model with different α (Female speech)

Male Speech 1

Input SNR (dB)	Input Signal LLR	Log Likelihood Ratio (LLR) Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	2.49	2.15	1.91	1.76	1.70	1.64
5	2.01	1.73	1.53	1.42	1.39	1.38
10	1.58	1.32	1.15	1.11	1.12	1.17
15	1.17	0.98	0.86	0.85	0.88	0.95
20	0.81	0.68	0.61	0.64	0.69	0.80

Male Speech 2

Input SNR (dB)	Input Signal LLR	Log Likelihood Ratio (LLR) Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	1.38	1.27	1.19	1.17	1.18	1.23
5	1.08	0.91	0.83	0.83	0.86	0.91
10	0.78	0.65	0.60	0.64	0.68	0.77
15	0.53	0.44	0.42	0.50	0.56	0.66
20	0.36	0.30	0.30	0.39	0.46	0.60

Male Speech 3

Input SNR (dB)	Input Signal LLR	Log Likelihood Ratio (LLR) Fixed α Subtraction				
		α=1	α=0.50	α=0.33	α=0.27	α=0.20
0	1.67	1.55	1.47	1.46	1.47	1.50
5	1.35	1.18	1.09	1.08	1.11	1.19
10	1.08	0.91	0.92	0.83	0.87	0.97
15	0.79	0.68	0.65	0.71	0.77	0.90
20	0.55	0.47	0.43	0.49	0.55	0.68

Table 4.6: Comparison of Log Likelihood Ratios for subtraction model with different α (Male speech)

consistent results. These test results lead us to propose an adaptive-α approach for speech enhancement where α is adapted on a frame-by-frame basis. When the SNR of the frame is low, a smaller α will be assigned to remove more noise. When the SNR of the frame is high, a larger α will be assigned to better preserve the speech. Note that the SNR changes rapidly from one frame to another under the noisy condition. Also the clean speech variance is not available. This leads us to consider an adaptive choice of α based on the estimated Noisy Signal to Noise Ratio (NSNR) estimated using the variance estimates λ_y and λ_n in equations (4.3) and (4.2).

4.3 Adaptive-α Subtraction Speech Enhancement Algorithm

In the previous section, we noted that the performance of speech enhancement depends on the choice of α. With higher input SNR, the value of α needs to be increased and vice versa. In this section, we will present a statistical approach for determining the value of α based on the loudness measure of speech.

4.3.1 Selection of α based on Statistical Modelling

First, we will try to determine the value of α based on minimizing the error in the loudness domain. As previously suggested, the Laplacian model is preferable over Gaussian model for speech signals [47]. Hence, stationary Laplacian and Gaussian sources are used for speech and noise, respectively.

In the simulation, a Gaussian source N (Noise) and a Laplacian source X (clean speech signal) with a certain ratio of SNR are used. The noisy signal is simulated as $Y = X + N$.

The loudness model of (2.60) is used here due to its simplicity. The problem is to find the best estimate $\hat{X}$ so that the error in loudness domain:

$$E_L = \left((X^2)^{\alpha_0} - \left(\hat{X}^2\right)^{\alpha_0} \right)^2 \tag{4.4}$$

is minimized with α_0 chosen as 0.27 as given in [58].

It should be noted that we are not discussing adapting α_0. Rather, the objective of the simulation is to find the value of α that produces the best estimate $\hat{X}$ to minimize E_L in (4.4):

$$\left(\hat{X}^2\right)^{\alpha} = (Y^2)^{\alpha} - (N^2)^{\alpha}. \tag{4.5}$$

The simulation is done using SNR from -10 dB to 30 dB and α is tested for values varying from 0 to 1 with 0.05 increment. The value of α that results in $\hat{X}$ that minimizes E_L in (4.4) is recorded and results are given in Figure 4.1. The SNR is transferred to the corresponding NSNR values for the X-axis.

From Figure 4.1, we can clearly see that the best value of α increases as NSNR increases. We can also see that the best value of α starts from 0.3 at -10 dB (0.4 dB NSNR) and approaches 1 around 12 dB. This is partly in agreement with the results obtained in the previous section. It shows that the best choice of α does increase with the input SNR. However, due to the fact that the actual SNR in a noisy speech signal varies in a much larger range than in this simple statistical model, the actual selection of the suitable α values is more complex.

4.3.2 Adaptive-α Subtraction Speech Enhancement

In the following simulation, different samples of male and female speech are used for testing. Speech signals are mixed with uncorrelated white noise and the overall SNR varies

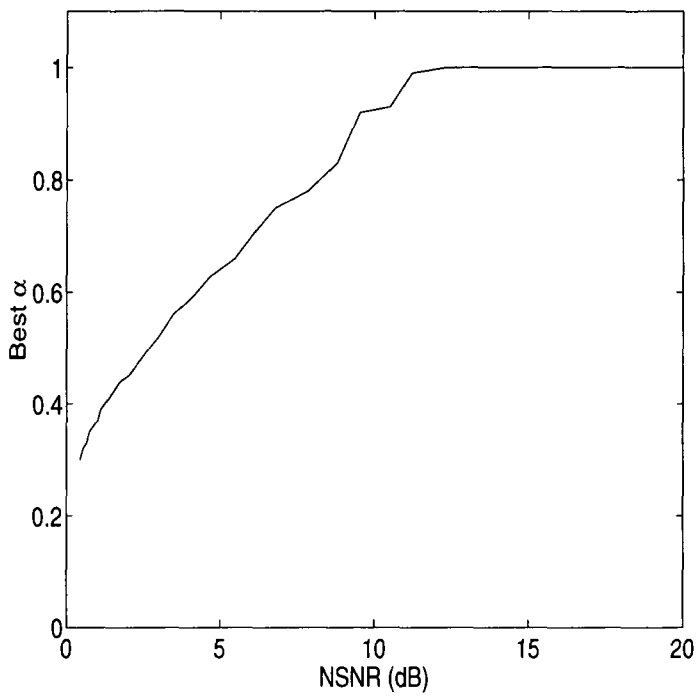

Figure 4.1: The best value of α for different NSNRs

from 0 dB to 20 dB with 5 dB increments. The noisy speech sample is preprocessed and transformed with FFT of length 128 with 50% overlap. The adaptive-α subtraction is then applied. Similar to the simulations in the previous chapter, the noise variance is estimated with a noise memory. The block diagram of such system is shown in Figure 4.2.

An simplified relationship between α and NSNR is given in Figure 4.3. The low end of α is selected as 0.3 below 0 dB NSNR and the high end of α stays at 1 for NSNR$\gg$ 12 dB. α changes linearly with the SNR between 0 and 12 dB. In the following Tables 4.7 and 4.8, this relationship is referred to as [0.3 1][0 12 dB].

Results for a set of choices for α are shown in Tables 4.7 to 4.8. The lower end of the α value is selected as 0.2 or 0.3 and the high end of the α is 0.5, 0.7, or 1. The value of α is changing linearly with the input NSNR for simplicity. The choice of the lower end and

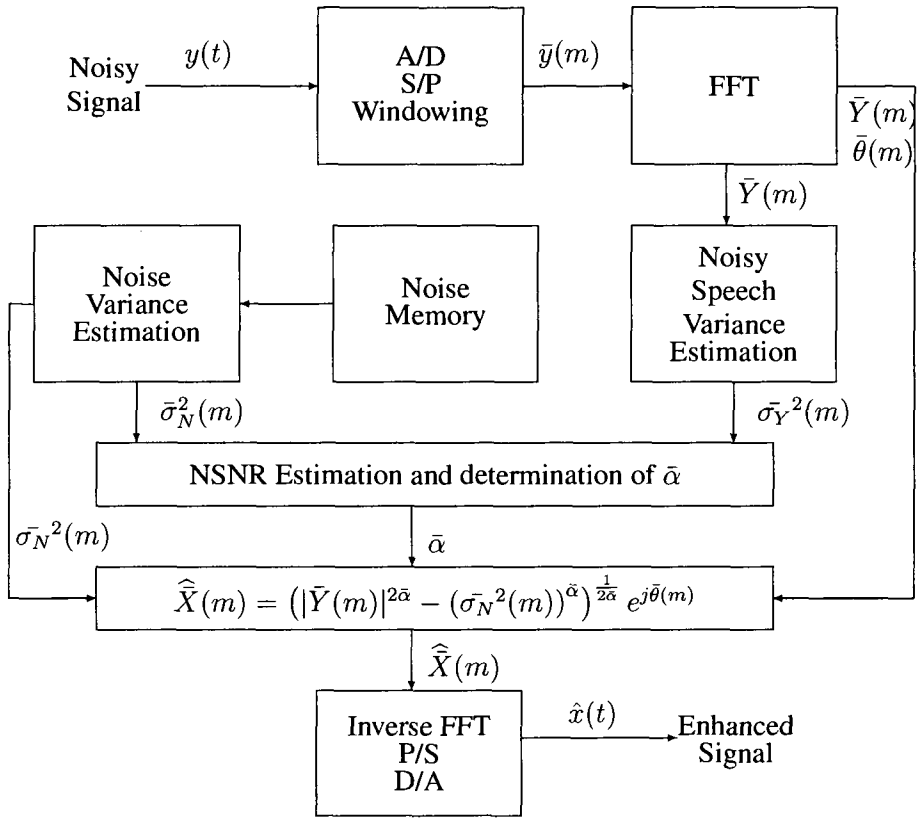

Figure 4.2: The block diagram of the adaptive α speech enhancement system

higher end values was chosen based on results from the simulations with fixed-α and the statistical modelling.

The PESQ, segmental SNR and LLR measures are evaluated separately. The first set of Tables 4.7 and 4.8 summarize the improvement in PESQ using different ranges for α for both male and females speech. [0.2 1][0 20dB] and [0.2 0.7][0 15dB] are the best two sets based on the overall performance. Especially, the set of [0.2 0.7][0 15dB] outperforms the best PESQ value achieved with fixed α's consistently. And the set of [0.2 1][0 20dB] performs better in 27 of 30 tests and just slightly worse in the others.

Tables 4.9 and 4.10 show the segmental SNRs of the algorithm. Similarly, the sets of

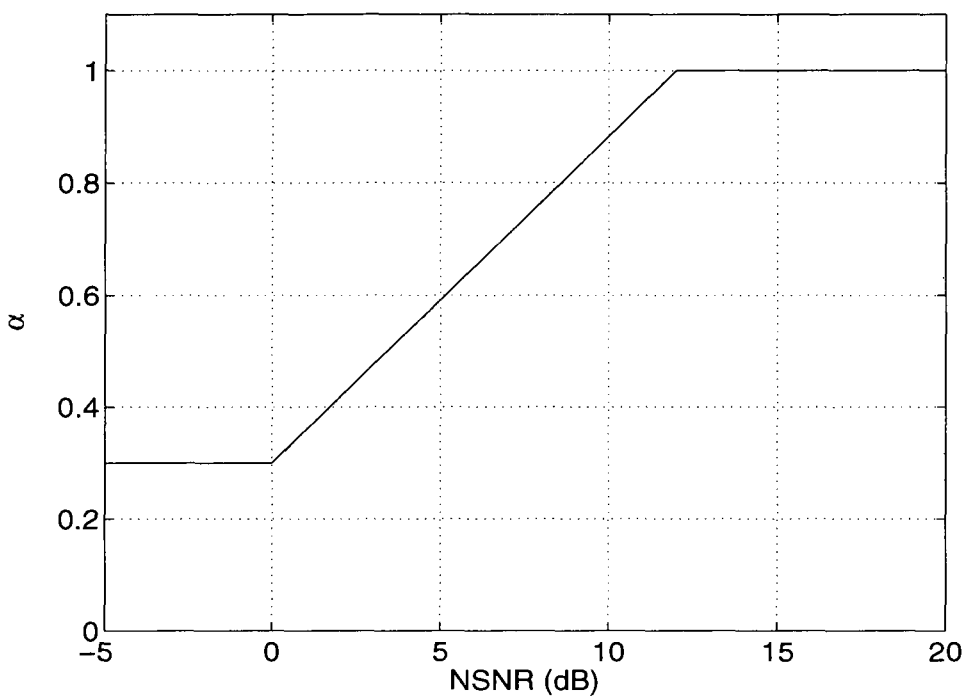

Figure 4.3: The relationship between SNR and power factor α

[0.2 1][0 20dB] and [0.2 0.7][0 15dB] provide the most consistent improvements. The set of [0.2 0.7][0 15dB] outperforms the best with fixed-α in all but one case.

The set of [0.3 1][0 12dB], which is suggested by the statistical modelling as in Figure 4.1, shows least SNR improvement when the input SNR is low. The residual noise is still high after enhancement. By further reducing the lower end of α to 0.2, the performance improves at lower input SNR cases.

Female Speech 1

Input SNR	Input PESQ	PESQ score improvements					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	1.36	+0.29	+0.38	+0.39	+0.40	+0.42	+0.37
5	1.64	+0.35	+0.44	+0.46	+0.47	+0.47	+0.42
10	1.93	+0.42	+0.51	+0.52	+0.52	+0.49	+0.44
15	2.32	+0.41	+0.51	+0.51	+0.50	+0.43	+0.37
20	2.67	+0.41	+0.51	+0.51	+0.48	+0.37	+0.36

Female Speech 2

Input SNR	Input PESQ	PESQ score improvements					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	1.60	+0.17	+0.20	+0.20	+0.19	+0.18	+0.17
5	1.87	+0.24	+0.29	+0.30	+0.31	+0.28	+0.20
10	2.17	+0.28	+0.31	+0.31	+0.29	+0.21	+0.21
15	2.44	+0.30	+0.34	+0.34	+0.34	+0.27	+0.24
20	2.73	+0.34	+0.38	+0.38	+0.37	+0.30	+0.25

Female Speech 2

Input SNR	Input PESQ	PESQ score improvements					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	1.25	+0.42	+0.55	+0.57	+0.60	+0.62	+0.55
5	1.60	+0.45	+0.54	+0.55	+0.56	+0.56	+0.51
10	1.96	+0.43	+0.54	+0.55	+0.55	+0.50	+0.45
15	2.33	+0.42	+0.53	+0.54	+0.53	+0.47	+0.44
20	2.70	+0.44	+0.52	+0.52	+0.50	+0.38	+0.35

Table 4.7: Comparison of PESQ scores with adaptive-α subtraction and different set of factors (Female speech)

Male Speech 1

Input SNR	Input PESQ	PESQ score improvements					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α=
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	1.39	+0.22	+0.30	+0.31	+0.32	+0.32	+0.15
5	1.57	+0.31	+0.41	+0.42	+0.43	+0.43	+0.35
10	1.83	+0.40	+0.48	+0.49	+0.48	+0.43	+0.38
15	2.15	+0.36	+0.41	+0.41	+0.39	+0.31	+0.33
20	2.49	+0.31	+0.32	+0.31	+0.27	+0.18	+0.27

Male Speech 2

Input SNR	Input PESQ	PESQ score improvements					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α=
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	1.63	+0.22	+0.29	+0.29	+0.31	+0.31	+0.29
5	1.95	+0.39	+0.46	+0.47	+0.48	+0.48	+0.38
10	2.29	+0.39	+0.50	+0.50	+0.50	+0.45	+0.44
15	2.68	+0.38	+0.44	+0.44	+0.43	+0.35	+0.37
20	3.11	+0.22	+0.30	+0.31	+0.29	+0.20	+0.25

Male Speech 3

Input SNR	Input PESQ	PESQ score improvements					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α=
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	1.35	+0.39	+0.51	+0.53	+0.55	+0.56	+0.52
5	1.70	+0.49	+0.62	+0.63	+0.66	+0.67	+0.60
10	2.06	+0.52	+0.61	+0.61	+0.60	+0.56	+0.51
15	2.47	+0.46	+0.55	+0.55	+0.54	+0.46	+0.45
20	2.84	+0.38	+0.50	+0.51	+0.51	+0.42	+0.37

Table 4.8: Comparison of PESQ scores with adaptive-α subtraction and different set of factors (Male speech)

Female Speech 1

Input SNR	Input PESQ	Segmental SNR					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	-6.56	0.19	2.08	2.36	2.60	2.49	1.51
5	-1.56	4.04	5.25	5.36	5.37	4.81	4.04
10	3.44	7.65	8.25	8.23	8.06	7.22	7.53
15	8.44	11.37	11.60	11.49	11.20	10.10	11.03
20	13.44	15.22	15.13	14.94	14.53	13.17	14.41

Female Speech 2

Input SNR	Input PESQ	Segmental SNR					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	-11.78	-4.79	-2.36	-1.83	-1.15	-0.38	-0.25
5	-6.78	-0.80	1.10	1.46	1.89	2.16	0.68
10	-1.78	3.10	4.38	4.58	4.76	4.59	3.21
15	3.22	6.90	7.68	7.74	7.73	7.15	6.76
20	8.22	10.76	11.21	11.19	11.04	10.13	10.37

Female Speech 3

Input SNR	Input PESQ	Segmental SNR					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	-6.76	-0.44	1.27	1.55	1.83	1.82	0.77
5	-1.76	3.48	4.59	4.72	4.75	4.30	3.59
10	3.24	7.47	8.20	8.23	8.12	7.32	7.08
15	8.24	11.11	11.53	11.51	11.34	10.38	10.77
20	13.24	15.13	15.17	15.03	14.72	13.45	14.44

Table 4.9: Comparison of Segmental SNR with adaptive-α subtraction and different set of factors (Female speech)

Male Speech 1

Input SNR	Input PESQ	Segmental SNR					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	-6.30	0.39	2.26	2.55	2.80	2.69	1.55
5	-1.30	4.43	5.86	6.04	6.12	5.58	4.35
10	3.70	8.31	9.19	9.24	9.15	8.33	8.03
15	8.70	11.92	12.40	12.35	12.16	11.15	11.64
20	13.70	15.74	15.79	15.65	15.37	14.14	15.18

Male Speech 2

Input SNR	Input PESQ	Segmental SNR					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	-5.95	-0.28	1.26	1.56	1.85	1.93	0.95
5	-0.95	3.64	4.66	4.81	4.88	4.56	3.74
10	4.05	7.27	7.84	7.88	7.81	7.19	7.41
15	9.05	11.23	11.39	11.31	11.04	10.00	10.90
20	14.05	15.29	15.30	15.17	14.84	13.53	14.51

Male Speech 3

Input SNR	Input PESQ	Segmental SNR					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	-7.87	-1.39	0.54	0.90	1.31	1.58	0.59
5	-2.87	2.49	3.78	3.96	4.12	3.87	2.71
10	2.13	6.08	6.88	6.95	6.91	6.34	6.13
15	7.13	9.85	10.20	10.14	9.90	8.95	9.51
20	12.13	13.83	13.90	13.77	13.45	12.22	13.18

Table 4.10: Comparison of Segmental SNR with adaptive-α subtraction and different set of factors (Male speech)

Female Speech 1

Input SNR	Input PESQ	LLR					
		Adaptive $[\alpha_{min}\ \alpha_{min}]\ [SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	2.02	1.58	1.55	1.55	1.55	1.56	1.68
5	1.66	1.27	1.25	1.25	1.26	1.28	1.29
10	1.29	0.90	0.90	0.91	0.93	0.95	1.00
15	0.91	0.62	0.64	0.65	0.67	0.71	0.76
20	0.60	0.42	0.46	0.48	0.50	0.56	0.52

Female Speech 2

Input SNR	Input PESQ	LLR					
		Adaptive $[\alpha_{min}\ \alpha_{min}]\ [SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	1.61	1.46	1.47	1.47	1.47	1.47	1.68
5	1.31	1.16	1.19	1.19	1.19	1.20	1.29
10	1.06	0.86	0.92	0.92	0.93	0.95	1.00
15	0.78	0.70	0.77	0.79	0.81	0.84	0.76
20	0.56	0.46	0.52	0.53	0.56	0.61	0.52

Female Speech 3

Input SNR	Input PESQ	LLR					
		Adaptive $[\alpha_{min}\ \alpha_{min}]\ [SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	2.12	1.61	1.51	1.50	1.48	1.46	1.53
5	1.72	1.18	1.15	1.14	1.14	1.13	1.17
10	1.31	0.89	0.88	0.88	0.90	0.94	0.98
15	0.96	0.64	0.65	0.66	0.69	0.74	0.68
20	0.64	0.46	0.47	0.49	0.52	0.59	0.49

Table 4.11: Comparison of Log Likelihood Ratio with adaptive-α subtraction and different set of factors (Female speech)

Male Speech 1

Input SNR	Input PESQ	LLR					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	2.49	1.79	1.64	1.63	1.63	1.61	1.64
5	2.01	1.44	1.34	1.33	1.29	1.27	1.38
10	1.58	1.09	1.06	1.06	1.07	1.08	1.11
15	1.17	0.83	0.82	0.82	0.83	0.86	0.85
20	0.81	0.59	0.59	0.60	0.61	0.65	0.61

Male Speech 2

Input SNR	Input PESQ	LLR					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	1.38	1.14	1.14	1.14	1.14	1.15	1.17
5	1.08	0.82	0.83	0.83	0.85	0.86	0.83
10	0.78	0.59	0.63	0.65	0.70	0.75	0.64
15	0.53	0.41	0.44	0.47	0.51	0.55	0.42
20	0.36	0.28	0.30	0.35	0.38	0.44	0.30

Male Speech 3

Input SNR	Input PESQ	LLR					
		Adaptive $[\alpha_{min}\ \alpha_{min}]$ $[SNR_{min}\ SNR_{min}]$					Best with fixed factor α
		[0.3 1] [0 12dB]	[0.2 1] [0 20dB]	[0.2 0.7] [0 15dB]	[0.2 0.7] [0 20dB]	[0.2 0.5] [0 20dB]	
0	1.67	1.45	1.47	1.48	1.48	1.49	1.46
5	1.35	1.08	1.10	1.13	1.15	1.17	1.08
10	1.08	0.81	0.84	0.86	0.88	0.92	0.83
15	0.79	0.64	0.65	0.68	0.74	0.80	0.65
20	0.55	0.42	0.43	0.44	0.47	0.53	0.43

Table 4.12: Comparison of Log Likelihood Ratio with adaptive-α subtraction and different set of factors (Male speech)

Tables 4.11 and 4.12 show the log likelihood ratio (LLR) measure of the enhanced speech. The results show that the two sets of [0.2 1][0 20dB] and [0.2 0.7][0 15dB] provide comparable LLR results with the best achieved with fixed-α.

Next, we examine the enhanced signal energy in Table 4.13. This table shows the energy of the enhanced signal for the adaptive-α subtraction, fixed-α subtraction compared with original signal.

Input SNR	Input Speech Variance	Enhanced Signal Variances				
		Adaptive	Fixed factor			
		[0.2 0.7][0 15dB]	0.20	0.27	0.33	0.50
0	0.14	0.13	0.02	0.06	0.08	0.12
5	0.14	0.14	0.04	0.07	0.09	0.12
10	0.14	0.14	0.05	0.09	0.10	0.13
15	0.14	0.14	0.06	0.10	0.11	0.13
20	0.14	0.14	0.08	0.11	0.12	0.14

Table 4.13: Comparison of enhanced signal variance with adaptive-α subtraction and different set of factors

It is clear that the fixed-α approach reduces the energy of the enhanced signal significantly when the α value is low, or when the input SNR is low. On the other hand, the adaptive-α subtraction can preserve the speech signal much better than the fixed-α case.

Overall, we expected to improve performance while adapting α to the best value for the frame-by-frame SNR. This has been confirmed by the simulations. Based on the results, the adaptive-α subtraction has the following advantages:

• The adaptive α will improve the performance of the speech enhancement over a fixed α. The set of [0.3 0.7][0 15dB] can provide better PESQ and segmental SNR improvements than the fixed α. With adaptive α, the choice of α is on a frame-by-frame basis. Unlike the fixed-α, no estimation of the overall SNR is required for the selection of the proper value of α. The algorithm adjusts the value adaptively using the information from the most recent

frame.

• The best set of adaptive α values is [0.2 0.7][0 15dB]. It provides consistent results for all the tests. It shows further improvements in the PESQ and segmental SNR measures and no serious distortion is introduced.

• With the fixed α, the energy will be significantly reduced for low SNR or smaller α cases. This is due to the severe over-subtraction for these conditions. The adaptive α will preserve the energy of the original speech as shown in Table 4.13 compared to the fixed α case which attenuated the signal at low SNR or smaller α cases.

4.4 Conclusions

In this Chapter, we presented an adaptive-α subtraction speech enhancement approach. This approach was originally based on the generalized spectral subtraction method. The value of α used in the generalized spectral subtraction field has been relatively subjective.

We first tested the performance of speech enhancement with fixed α for all frames in a speech file. We tested for various values of α and evaluate the performance based on three objective quality measures: PESQ, segmental SNR and log likelihood ratio. It was shown that the best α depends on the SNR. This leads us to propose the adaptation of α based on frame-by-frame SNRs.

We also proposed a statistical method to choose the range of the values of α as a function of the SNR. We used a stationary Gaussian source for noise and a Laplacian source for speech in the simulation. The α value selected minimizes the error in the loudness domain. The results confirmed that the performance improves when the value of α is chosen adaptively based on the frame-by-frame SNR, with α increasing for higher SNR.

Next, we used the results of the statistical test to enhance speech with the adaptive-α. We used several sets of different α values for testing, which were chosen based on the results of the fixed-α and the statistical modelling. We showed that the adaptive-α subtraction speech enhancement algorithm can improve the performance consistently over fixed α. The adaptive approach was shown to be superior as assessed by the PESQ and segmental SNR measures. The LLR measure confirmed that the adaptive approach does not lead to further distortion. We also noted that adaptive-α can preserve the energy of the speech in the original noisy signals. All simulation results confirm that the adaptation of α in a loudness-based speech enhancement algorithm is an efficient method that improves the performance of the generalized spectral subtraction speech enhancement algorithm.

Chapter 5

Speech Enhancement based on Loudness Subtraction Model

5.1 Introduction

Spectral subtraction has been used as an computationally effective method for speech enhancement [26] [80]. This approach estimates the power spectral density of the clean speech signal by subtracting the power spectral density (PSD) of the noise estimate from the PSD of the noisy signal. Also the magnitude estimates [26] and the envelope of the spectrum [106] have been used as substitutes for PSD in the subtraction .

Even though the additive noise is usually assumed to be stationary, it still has its peaks and valleys for a particular frame. Simple spectral subtraction may lead to unpleasant residue in the enhanced speech signals. In [25], an over-subtraction approach was proposed. The over-subtraction subtracts an over-estimate by adding a multiplication factor before the noise estimate in the subtraction process. It eliminates the noise peaks better compared to the normal subtraction. The overall performance can be improved by carefully selecting the

parameters for the over-subtraction. The parameters are selected based on the estimation of the instantaneous SNR. When the instantaneous SNR is low, the parameters will be selected to remove more noise, and vice versa. This over-subtraction approach has also been used in the substitutional algorithms for magnitude [26] or spectrum envelope [106].

The spectral subtraction has been further improved with the consideration of the masking property of the human ear [109]. The noise masking threshold, below which the noise is inaudible to the human ear, is calculated. The subtraction adjusts the level of the subtraction based on the noise masking threshold. When the threshold is low, more noise is removed from the noisy speech, and vice versa. Thus, the subtraction factor is controlled by the noise masking threshold.

In searching for a speech enhancement algorithm with better performance, we note that the perceived quality of the enhanced speech is eventually measured by Mean Opinion Score (MOS). It was shown in [95] that MOS can be measured with reasonable accuracy by the ITU-T Recommendation P.862 (PESQ). As such, PESQ has been used to estimate the performance of a speech enhancement algorithm. The PESQ uses the loudness model to compare a (distorted) speech signal to a reference (clean) speech signal. The loudness density differences between the distorted and clean signals are considered as the disturbance. This disturbance is the main measure of the speech quality by PESQ. This leads us to try developing speech enhancement algorithms to minimize the loudness disturbance. As mentioned before, the loudness model is a more accurate model for the human hearing system and as such, we expect a loudness-based subtraction-type speech enhancement algorithm can outperform the corresponding algorithms in the spectral domain.

In this chapter, the loudness subtraction and over-subtraction algorithms are proposed.

5.2 Direct Loudness Substraction

The speech enhancement algorithm using direct loudness subtraction was originally proposed by Petersen [93]. The loudness of a signal with intensity I is:

$$L_X = C \cdot \left(\left(X^2 + M \right)^{\alpha} - (M)^{\alpha} \right). \tag{5.1}$$

The clean speech signal spectrum is X, C is a constant and M is the internal masking intensity, which is 4 dB above the absolute threshold [93]. The absolute threshold is the minimum sound level of a pure tone that an average ear with normal hearing can hear in a noiseless environment [10]. We assume the added noise energy is N^2. When the external noise is added to the speech, the human ear combines it with the internal masking noise as the new masking noise [82]. The loudness of the resulting signal perceived by the ear can be formulated as

$$L_X = C \cdot \left(\left(X^2 + N^2 + M \right)^{\alpha} - \left(N^2 + M \right)^{\alpha} \right). \tag{5.2}$$

Assume the enhanced signal with energy $\hat{X}^2$ has the same loudness as the above perceived loudness and has only the internal masking effect. Then,

$$C \cdot \left(\left(\hat{X}^2 + M \right)^{\alpha} - (M)^{\alpha} \right) = C \cdot \left(\left(X^2 + N^2 + M \right)^{\alpha} - \left(N^2 + M \right)^{\alpha} \right). \tag{5.3}$$

Thus, the energy of the enhanced signal, with the same loudness as the noisy signal, is given by;

$$\hat{X}^2 = \left(\left(Y^2 + M \right)^{\alpha} - \left(N^2 + M \right)^{\alpha} + M^{\alpha} \right)^{\frac{1}{\alpha}} - M. \tag{5.4}$$

where $Y^2 = X^2 + N^2$ is the energy of the noisy signal.

Overall, this method is based on the loudness subtraction approach in the frequency domain. This approach was shown to provide smoother and less-distorted speech than spectral subtraction [93].

5.3 Statistical Loudness Subtraction Model

In this section, the loudness subtraction model of speech enhancement is considered. The human hearing system assesses the relative quality of a speech signal based on the difference in the loudness domain between the given signal and the reference (clean) speech signal. This suggests a subtraction-based speech enhancement in the loudness domain.

In [93], a speech enhancement algorithm based on the loudness subtraction with considerations of the internal masking intensity. This internal masking intensity represents the level of a tone to be just audible. When the signal loudness is high, the effect of the internal masking is negligible. Thus, we can rewrite (5.4) as:

$$\hat{X} = \left((Y)^{2\alpha} - (N)^{2\alpha}\right)^{1/2\alpha}. \tag{5.5}$$

In general, for $\alpha \neq 1$, and even if X and N are assumed to be independent, the following equation holds:

$$E\left\{Y^{2\alpha}\right\} \neq E\left\{X^{2\alpha}\right\} + E\left\{N^{2\alpha}\right\}. \tag{5.6}$$

Considering both (5.5) and (5.6), the average loudness of the enhanced signal will not be automatically equal to the average loudness of original signal: $E\left\{\hat{X}^{2\alpha}\right\} \neq E\left\{X^{2\alpha}\right\}$, even if the noise signal is accurately estimated and subtracted out.

A more generalized loudness subtraction algorithm is proposed as follows:

$$\hat{X}^2 = \left(\left(Y^2\right)^{\alpha} - a\left(N^2\right)^{\alpha}\right)^{\frac{1}{\alpha}} \tag{5.7}$$

where a is defined as the subtraction factor. To maintain the loudness of the reconstructed speech at the same levels as the original clean speech, we require

$$E\left\{\hat{X}^{2\alpha}\right\} = E\left\{Y^{2\alpha}\right\} - E\left\{aN^{2\alpha}\right\} \cong E\left\{X^{2\alpha}\right\} \tag{5.8}$$

where

$$a \cong \frac{E\{Y^{2\alpha} - X^{2\alpha}\}}{E\{N^{2\alpha}\}} \tag{5.9}$$

Defining the Noisy-Signal-to-Noise Ratio (NSNR) as:

$$NSNR = \frac{E\{Y^2\}}{E\{N^2\}}, \tag{5.10}$$

we now proceed to determine a in terms of measurable signals.

In the frequency domain, the clean speech X is assumed to have Laplacian distribution, while the noise N is assumed to be Gaussian. To simulate this, two stationary sources (Laplacian and Gaussian) are used as samples for X and N in the above equations. A Laplacian source can be obtained using a function of a uniform distributed source U in $[0, 1]$ with the following transform:

$$X = \log(1 - U). \tag{5.11}$$

Random selected phases with uniform distribution between 0 and 2π are added to X and N, respectively. Both sources are zero mean and with variances determined by the specified NSNR.

Figure 5.1 shows the parameter a as a function of the NSNR calculated as in (5.9). The NSNR is estimated on a frame-by-frame basis to adjust the value of a with the variance of the speech signals. This will be needed later in the speech enhancement algorithms.

5.3.1 Simulation Setup

In the next section, several proposed speech enhancement approaches will be evaluated using the ITU P.862 (PESQ) [15].

White noise is added to the clean speech signal with different SNRs from 0 dB to 20 dB. The noisy speech goes through a serial to parallel converter to be separated into frames

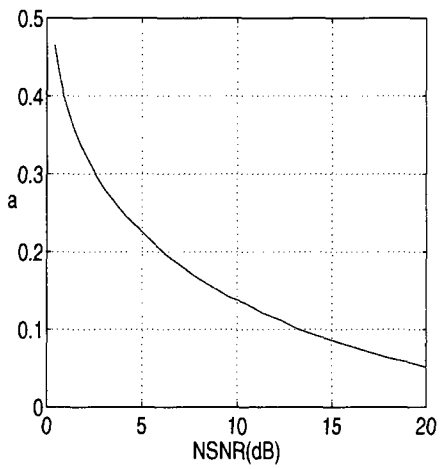

Figure 5.1: Statistical estimation of the subtraction factor a for loudness subtraction (Eq. (5.9))

of 128 samples each. Then the FFT of each frame is obtained resulting in the noisy speech spectral magnitude vector $\bar{Y}$. The phase of the FFT is saved separately.

The loudness of the noise is estimated separately using a noise memory [46] [48] [118]. In practical systems, the noise memory can be derived from the noisy-only intervals from the noisy speech, which is detected by a voice activity detector (VAD). The noise loudness is estimated from the most recent noise samples, which have the same statistical property as the noise added to the speech. The use of noise memory can provide an excellent estimate of the noise loudness.

The FFT amplitude components for the enhanced signal are estimated using the different speech enhancement algorithms to be presented. The phase of the FFT of noisy speech is maintained for the enhanced speech. The Inverse FFT is applied to derive the enhanced speech in the corresponding frame. Finally, the frame-by-frame speech signals is passed through a parallel to serial converter and the enhanced speech signal is reconstructed.

5.3.2 Proposed Loudness Subtraction with Fixed Subtraction Factor a

Here, we consider the algorithm proposed in Eq. (5.7) where a is estimated as in Figure 5.1. Speech signals are mixed with uncorrelated white noise and the overall SNR varies from 0 dB to 20 dB with 5 dB increments. Different samples of male and female speech are used for testing. These speech files will be used through the remainder of this chapter.

The noisy speech sample is preprocessed and transformed with FFT of length 128 with 50% overlap. Loudness subtraction is then applied using a fixed a for all frames for the whole speech within each test. A noise reference is used to generate the estimate of noise loudness. In this thesis, a white noise memory with the same variance as the additive noise is used for testing the speech enhancement algorithm. More complicated noise estimation methods for non-stationary noise have been used in speech enhancement. In this thesis, we only test with stationary noise and the reference memory can provide a good estimation for the noise loudness. Under such situation, the result can reflect the performance of speech enhancement algorithm itself. It leads to easier comparison with other algorithms.

The block diagram of such system is shown in Figure 5.2.

The objective of this test is to identify the optimal value of a for different overall SNRs based on the PESQ quality measure. Eight different values of a from 0.2 to 2 are tested with speech signals. The results are given in Tables 5.1 to 5.2.

The results from the Tables 5.1 to 5.2 show that the best choice of a decreases as the SNR increases. This is consistent with the statistical estimation (Figure 5.1) in the previous section. But the optimal value of a is much higher than the value as in Figure 5.1. The optimal value ranges from 0.6 to 1.2, compared to the range of 0.05 to 0.5 in the statistical model as shown in Figure 5.1.

Female Speech 1

Input SNR (dB)	Input PESQ	PESQ score improvements							
		a=2	a=1.5	a=1.2	a=1	a=0.8	a=0.6	a=0.4	a=0.2
0	1.36	+0.01	+0.41	**+0.42**	+0.31	+0.25	+0.20	+0.14	+0.08
5	1.64	-0.01	+0.22	**+0.42**	+0.38	+0.32	+0.23	+0.14	+0.07
10	1.93	-0.16	+0.10	+0.39	**+0.44**	+0.39	+0.30	+0.20	+0.10
15	2.32	-0.38	-0.10	+0.30	+0.32	**+0.33**	+0.28	+0.20	+0.10
20	2.67	-0.53	-0.16	+0.07	+0.23	**+0.28**	+0.27	+0.20	+0.11

Female Speech 2

Input SNR (dB)	Input PESQ	PESQ score improvements							
		a=2	a=1.5	a=1.2	a=1	a=0.8	a=0.6	a=0.4	a=0.2
0	1.60	-0.55	-0.01	**+0.19**	+0.18	+0.16	+0.12	+0.10	+0.05
5	1.87	-0.60	-0.18	+0.21	**+0.21**	+0.20	+0.17	+0.12	+0.07
10	2.17	-0.66	-0.37	+0.03	+0.14	**+0.16**	+0.14	+0.12	+0.07
15	2.44	-0.81	-0.38	-0.04	+0.12	**+0.19**	+0.19	+0.11	+0.06
20	2.73	-0.84	-0.36	-0.02	+0.17	**+0.19**	+0.17	+0.12	+0.06

Female Speech 3

Input SNR (dB)	Input PESQ	PESQ score improvements							
		a=2	a=1.5	a=1.2	a=1	a=0.8	a=0.6	a=0.4	a=0.2
0	1.25	-0.19	+0.24	**+0.53**	+0.49	+0.40	0.31	+0.18	+0.07
5	1.60	-0.35	+0.12	+0.45	**+0.46**	+0.42	+0.33	+0.23	+0.11
10	1.96	-0.36	+0.05	+0.33	**+0.44**	+0.42	+0.35	+0.25	+0.14
15	2.33	-0.50	-0.10	+0.25	+0.35	**+0.38**	+0.34	+0.23	+0.11
20	2.70	-0.59	-0.21	+0.07	+0.24	**+0.30**	+0.28	+0.24	+0.14

Table 5.1: PESQ scores with fixed a loudness subtraction (Female speech)

Male Speech 1

Input SNR (dB)	Input PESQ	PESQ score improvements							
		a=2	a=1.5	a=1.2	a=1	a=0.8	a=0.6	a=0.4	a=0.2
0	1.39	-0.02	+0.18	**+0.28**	+0.18	+0.12	+0.08	+0.08	+0.05
5	1.57	-0.07	+0.11	**+0.37**	+0.32	+0.24	+0.17	+0.10	+0.05
10	1.83	-0.29	-0.02	+0.31	**+0.38**	+0.34	+0.25	+0.16	+0.08
15	2.15	-0.55	-0.28	+0.12	+0.27	**+0.30**	+0.26	+0.20	+0.10
20	2.49	-0.68	-0.42	-0.10	+0.09	+0.20	**+0.23**	+0.18	+0.10

Male Speech 2

Input SNR (dB)	Input PESQ	PESQ score improvements							
		a=2	a=1.5	a=1.2	a=1	a=0.8	a=0.6	a=0.4	a=0.2
0	1.63	-0.41	+0.10	**+0.29**	+0.27	+0.23	+0.17	+0.11	+0.04
5	1.95	-0.23	+0.12	**+0.37**	+0.36	+0.34	+0.29	+0.15	+0.07
10	2.29	-0.48	-0.07	+0.30	**+0.37**	+0.35	+0.29	+0.25	+0.14
15	2.68	-0.65	-0.22	+0.18	+0.32	**+0.38**	+0.36	+0.25	+0.12
20	3.11	-0.80	-0.44	-0.13	+0.07	**+0.27**	+0.27	+0.22	+0.08

Male Speech 3

Input SNR (dB)	Input PESQ	PESQ score improvements							
		a=2	a=1.5	a=1.2	a=1	a=0.8	a=0.6	a=0.4	a=0.2
0	1.35	-0.01	+0.40	**+0.52**	+0.44	+0.37	+0.28	+0.18	+0.10
5	1.70	-0.12	+0.28	**+0.58**	+0.53	+0.43	+0.31	+0.20	+0.08
10	2.06	-0.26	+0.13	+0.41	**+0.52**	+0.46	+0.36	+0.22	+0.11
15	2.47	-0.45	-0.12	+0.28	+0.39	**+0.41**	+0.36	+0.21	+0.10
20	2.84	-0.58	-0.19	+0.067	+0.25	**+0.32**	+0.30	+0.23	+0.11

Table 5.2: PESQ scores with fixed a loudness subtraction (Male speech)

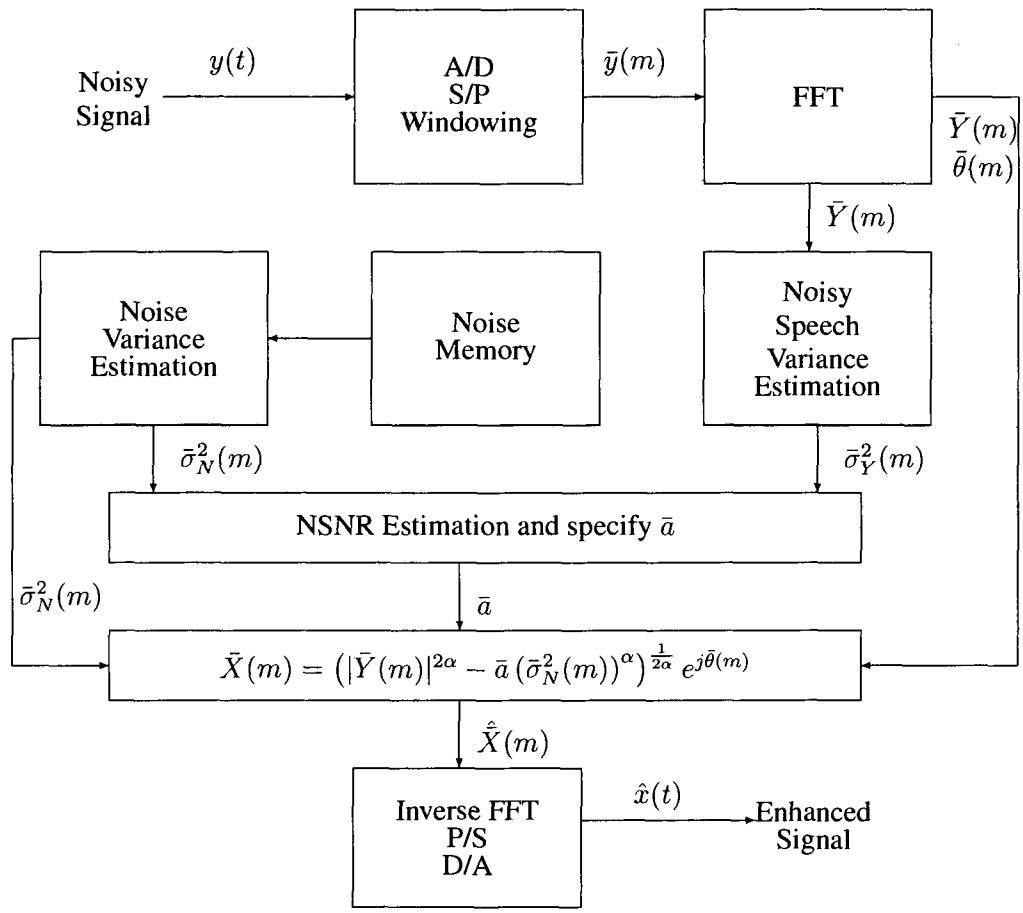

Figure 5.2: The block diagram of the loudness subtraction type speech enhancement system

The approach in the previous direct loudness subtraction in Eq. (5.5) subtracts the *average* loudness of the noise from the noisy speech signal. Given that the noise follows a Gaussian distribution, the *actual* loudness of the noise will differ from the average value for different frames. When the noise in a particular frame is large, there will be strong noise residue in the enhanced signal. When the noise is low, the fixed subtraction will lead to distortions to the speech.

Based on the above results, the effect of the value of a can be explained based on Eq.

(5.7). When a increases, the algorithm removes more noise with the associated higher chances of distortion of the speech itself. Such trade-off between distortion and noise deduction exists in many speech enhancement algorithms. If a is selected carefully, large noise residues can be removed only when necessary.

These conclusions lead us to the over-subtraction model that will be discussed next.

5.4 Proposed Loudness Over-Subtraction Model

5.4.1 Spectral Over-Subtraction

A spectral over-subtraction method has been proposed in [25] [107]. The main objective of the over-subtraction is to remove the peaks of the noise spectrum by removing an over-estimate of the noise. Especially when the SNR is low, a larger subtraction scaling factor was determined to be beneficial. The subtraction scaling factor for spectral over-subtraction [25] is shown as in Figure 5.3, with the corresponding spectral over-subtraction given by:

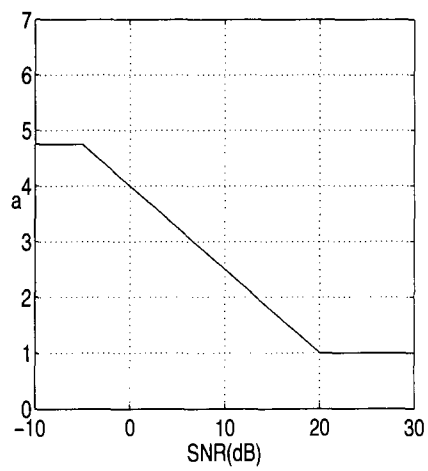

Figure 5.3: The subtraction factor a for spectral over-subtraction

$$\hat{X}^2 = Y^2 - E\{aN^2\}. \tag{5.12}$$

The enhanced speech signals of the spectral subtraction and spectral over-subtraction are shown in Figure 5.4.

Clearly, the spectral over-subtraction can remove more noise from the noisy speech signal, especially when the SNR is low. The speech portion is also lowered with over-subtraction. But the overall SNR is improved. The spectral over-subtraction can remove the large noise residues which can't be removed by the spectral subtraction. This leads us to propose a similar over-subtraction speech enhancement approach in the loudness domain next.

5.4.2 Loudness Over-Subtraction

In this section, we focus on removing a carefully determined over-estimate of the noise from the noisy signal in the loudness domain while avoiding increased distortion. As a is increased, more noise will be deducted but the enhanced speech will likely be more distorted. If a is chosen carefully, the improvement of the NSNR could compensate for the larger distortion of the speech. One natural way to modify the values in statistical loudness subtraction is through a scaling factor to a.

Two types of over-subtraction will be considered here:

- multiplying the values in statistical loudness subtraction by a fixed scaling factor.
- multiplying the values in statistical loudness subtraction by a linear scaling factor that has the shape as in Figure 5.3 (loudness over-subtraction).

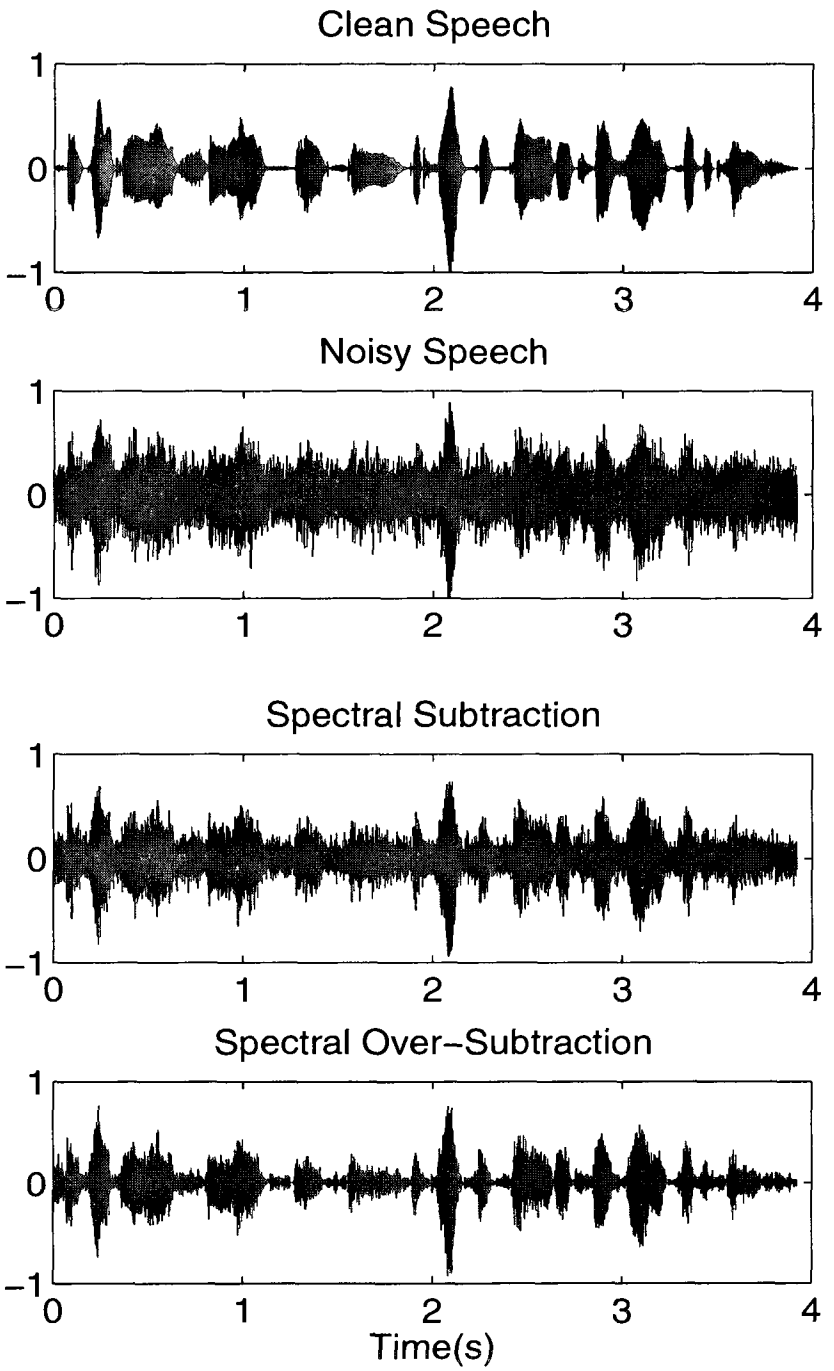

Figure 5.4: Comparison of the spectral over-subtraction with spectral subtraction, the vertical axis is the amplitude of speech signals

5.4.3 Fixed Multiplier Over-Subtraction

In this section, the a parameter in the statistical loudness subtraction model in Figure 5.1 will be multiplied by a factor c to generate a new $a_{multi} = ca$ value for speech enhancement. The term "scaling factor" will be used for this variable c. The constant c is tested with values of 2, 3, 4, 5, 6 and 7. These values are chosen based on the discussions in Section 5.3.2. The results are shown in Tables 5.3 to 5.4.

Female Speech 1

Input SNR (dB)	Input PESQ	PESQ score improvements					
		c=2	c=3	c=4	c=5	c=6	c=7
0	1.36	+0.26	+0.29	+0.34	**+0.39**	+0.38	+0.35
5	1.64	+0.32	+0.40	+0.466	**+0.51**	+0.44	+0.35
10	1.93	+0.39	+0.45	**+0.48**	+0.48	+0.43	+0.31
15	2.32	+0.35	+0.42	**+0.43**	+0.37	+0.30	+0.19
20	2.67	+0.39	**+0.43**	+0.41	+0.32	+0.28	+0.19

Female Speech 2

Input SNR (dB)	Input PESQ	PESQ score improvements					
		c=2	c=3	c=4	c=5	c=6	c=7
0	1.60	+0.09	+0.19	+0.22	**+0.23**	+0.11	+0.02
5	1.87	+0.17	+0.22	**+0.24**	+0.22	+0.01	-0.08
10	2.17	+0.22	**+0.22**	+0.22	+0.18	+0.04	-0.10
15	2.44	+0.25	**+0.27**	+0.22	+0.15	+0.14	+0.01
20	2.73	+0.25	+0.30	**+0.32**	+0.29	+0.17	+0.03

Female Speech 3

Input SNR (dB)	Input PESQ	PESQ score improvements					
		c=2	c=3	c=4	c=5	c=6	c=7
0	1.25	+0.32	+0.40	+0.44	+0.44	**+0.48**	+0.42
5	1.60	+0.41	+0.50	**+0.54**	+0.52	+0.48	+0.38
10	1.96	+0.39	+0.53	**+0.55**	+0.50	+0.41	+0.30
15	2.33	+0.38	+0.46	**+0.49**	+0.43	+0.34	+0.23
20	2.70	**+0.44**	+0.43	+0.40	+0.32	+0.30	+0.19

Table 5.3: PESQ scores with fixed multiplier loudness over-subtraction (Female speeches)

From Tables 5.3 to 5.4, the quality of enhanced speech can be further improved with

Male Speech 1

Input SNR (dB)	Input PESQ	PESQ score improvements					
		c=2	c=3	c=4	c=5	c=6	c=7
0	1.39	+0.13	+0.14	+0.19	+0.22	+0.27	**+0.27**
5	1.57	+0.27	+0.27	+0.32	+0.33	**+0.41**	+0.35
10	1.83	+0.35	+0.40	**+0.44**	+0.41	+0.36	+0.18
15	2.15	+0.32	**+0.38**	+0.36	+0.29	+0.19	+0.04
20	2.49	+0.27	**+0.29**	+0.26	+0.19	+0.05	-0.08

Male Speech 2

Input SNR (dB)	Input PESQ	PESQ score improvements					
		c=2	c=3	c=4	c=5	c=6	c=7
0	1.63	+0.25	+0.28	+0.30	**+0.31**	+0.29	+0.27
5	1.95	+0.27	+0.39	**+0.42**	+0.41	+0.29	+0.25
10	2.29	+0.41	**+0.45**	+0.43	+0.37	+0.33	+0.22
15	2.68	+0.40	**+0.44**	+0.43	+0.33	+0.17	+0.06
20	3.11	**+0.31**	+0.22	+0.15	+0.09	+0.08	-0.04

Male Speech 3

Input SNR (dB)	Input PESQ	PESQ score improvements					
		c=2	c=3	c=4	c=5	c=6	c=7
0	1.35	+0.24	+0.41	+0.49	**+0.53**	+0.49	+0.49
5	1.70	+0.41	+0.45	+0.51	**+0.53**	+0.51	+0.43
10	2.06	+0.43	+0.54	**+0.58**	+0.55	+0.46	+0.36
15	2.47	+0.37	**+0.46**	+0.46	+0.42	+0.35	+0.25
20	2.84	+0.36	**+0.43**	+0.41	+0.32	+0.23	+0.14

Table 5.4: PESQ scores with fixed multiplier loudness over-subtraction (Male speeches)

adding a multiply factor c to the statistical estimation of a. The best choice of the scale factor, c, varies with the input signal SNR. Overall, the best quality of enhanced signal is achieved at a scale factor of 2 or 3 for 20 dB SNR and 5 or 6 for 0 dB SNR.

5.4.4 Scaling Multiplier Loudness Over-Subtraction

In this section, the scaling factor will be determined based on the SNR of the speech frames. For simplicity, the scaling factor is assumed to be a linear function of the value of NSNR (in dB). Then the statistical loudness subtraction factor a is multiplied by the scaling factor to give the new over-subtraction factor.

Figure 5.5 shows one scaling factor choice as a function of NSNR. This over-subtraction scaling factor is chosen to be 4 below 0 dB NSNR, 1 beyond 20 dB NSNR and changing linearly between 0 and 20 dB. This set of scaling factors will be noted as $[5, 1]@[0, 20]$dB.

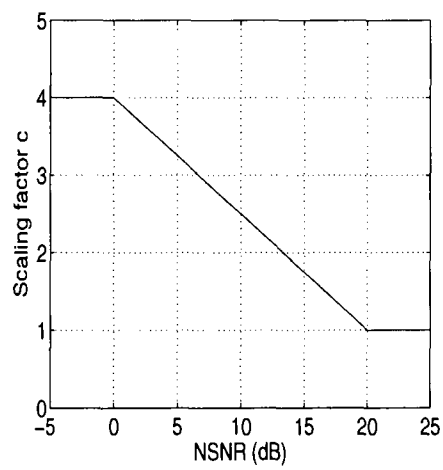

Figure 5.5: The scaling factor c for loudness over-subtraction

In this section, several sets of scaling factor are chosen for testing. The reason for

choosing these sets is based on the simulation results in the fixed multiplier loudness over-subtraction.

Female Speech 1

Input SNR (dB)	Input PESQ	PESQ score improvements					
		[7,1] @[0,20] dB	[6,1] @[0,20] dB	[6,2] @[0,20] dB	[5,1] @[0,20] dB	[5,2] @[0,20] dB	[4,1] @[0,20] dB
0	1.36	+0.44	+0.47	**+0.48**	+0.43	+0.42	+0.40
5	1.64	+0.46	+0.54	+0.52	**+0.55**	+0.47	+0.50
10	1.93	+0.38	+0.44	+0.42	+0.51	+0.50	**+0.54**
15	2.32	+0.32	+0.36	+0.33	+0.43	+0.44	**+0.50**
20	2.67	+0.36	+0.39	+0.36	+0.46	+0.44	**+0.51**

Female Speech 2

Input SNR (dB)	Input PESQ	PESQ score improvements					
		[7,1] @[0,20] dB	[6,1] @[0,20] dB	[6,2] @[0,20] dB	[5,1] @[0,20] dB	[5,2] @[0,20] dB	[4,1] @[0,20] dB
0	1.60	+0.09	+0.26	+0.24	+0.28	+0.16	**+0.33**
5	1.87	+0.05	+0.17	+0.15	+0.25	+0.18	**+0.30**
10	2.17	+0.09	+0.05	+0.02	+0.17	+0.23	**+0.32**
15	2.44	+0.14	+0.27	+0.24	+0.35	+0.30	**+0.37**
20	2.73	+0.25	+0.28	+0.28	+0.35	+0.34	**+0.39**

Female Speech 3

Input SNR (dB)	Input PESQ	PESQ score improvements					
		[7,1] @[0,20] dB	[6,1] @[0,20] dB	[6,2] @[0,20] dB	[5,1] @[0,20] dB	[5,2] @[0,20] dB	[4,1] @[0,20] dB
0	1.25	+0.56	+0.53	+0.53	+0.53	**+0.57**	+0.53
5	1.60	+0.46	+0.53	+0.50	+0.60	+0.57	**+0.61**
10	1.96	+0.41	+0.49	+0.47	+0.53	+0.51	**+0.56**
15	2.33	+0.33	+0.46	+0.44	+0.52	+0.47	**+0.53**
20	2.70	+0.33	+0.40	+0.38	+0.45	+0.42	**+0.53**

Table 5.5: PESQ scores with scaling multiplier type loudness over-subtraction (Female speech)

This scaling factor c is multiplied by a selected in the last section using (5.9) before being used in (5.7). With this scaling factor, the over-subtraction approach removes more noise when the NSNR is low and keeps the distortion low when the NSNR is high. The results are shown in Tables 5.5 to 5.6.

The results of Tables 5.5 to 5.6 are compared with Tables 5.3 to 5.4. The scaling

Male Speech 1

Input SNR (dB)	Input PESQ	PESQ score improvements					
		[7,1] @[0,20] dB	[6,1] @[0,20] dB	[6,2] @[0,20] dB	[5,1] @[0,20] dB	[5,2] @[0,20] dB	[4,1] @[0,20] dB
0	1.39	+0.29	+0.28	+0.26	+0.29	**+0.31**	+0.27
5	1.57	+0.32	+0.38	+0.35	+0.40	+0.38	**+0.43**
10	1.83	+0.32	+0.39	+0.35	**+0.47**	+0.41	+0.46
15	2.15	+0.20	+0.26	+0.23	+0.36	+0.36	**+0.42**
20	2.49	+0.15	+0.24	+0.21	+0.29	+0.26	**+0.32**

Male Speech 2

Input SNR (dB)	Input PESQ	PESQ score improvements					
		[7,1] @[0,20] dB	[6,1] @[0,20] dB	[6,2] @[0,20] dB	[5,1] @[0,20] dB	[5,2] @[0,20] dB	[4,1] @[0,20] dB
0	1.63	+0.32	+0.33	+0.32	+0.35	**+0.38**	+0.35
5	1.95	+0.48	+0.40	+0.39	+0.43	**+0.55**	+0.53
10	2.29	+0.38	+0.50	+0.48	+0.54	+0.48	**+0.54**
15	2.68	+0.30	+0.35	+0.33	+0.41	+0.39	**+0.50**
20	3.11	+0.18	+0.15	+0.13	+0.22	+0.27	**+0.33**

Male Speech 3

Input SNR (dB)	Input PESQ	PESQ score improvements					
		[7,1] @[0,20] dB	[6,1] @[0,20] dB	[6,2] @[0,20] dB	[5,1] @[0,20] dB	[5,2] @[0,20] dB	[4,1] @[0,20] dB
0	1.35	+0.56	+0.58	+0.60	+0.55	**+0.61**	+0.60
5	1.70	+0.53	+0.60	+0.58	+0.62	+0.55	**+0.62**
10	2.06	+0.49	+0.52	+0.49	+0.58	+0.58	**+0.64**
15	2.47	+0.39	+0.44	+0.42	+0.50	+0.47	**+0.58**
20	2.84	+0.30	+0.37	+0.35	+0.43	+0.38	**+0.49**

Table 5.6: PESQ scores with scaling multiplier type loudness over-subtraction (Male speech)

multiplier over-subtraction with the scaling factor [4,1]@[0,20] dB is shown to outperform or be comparable to the best possible results with fixed multiplier over-subtraction. Thus, we will use this linear scaling factor for our loudness over-subtraction approach.

With this approach, the scaling multiplier is variable as a function of the NSNR. The overall performance of speech enhancement is considerably better than any of the simple multiply type approach with one fixed multiply factor. Next, this approach will be compared with classical subtraction-type speech enhancement algorithms.

5.5 Comparison of the Subtraction-type Speech Enhancement Algorithms

5.5.1 Simulation results

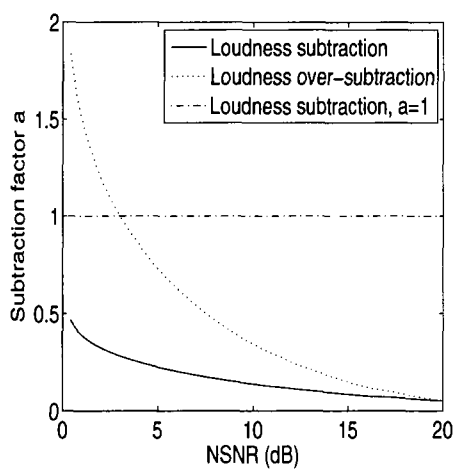

Figure 5.6: The subtraction factor for the approaches in loudness domain

In this section, the subtraction and over-subtraction approaches are compared in both

loudness and spectral domains. The results are evaluated using the ITU P.862 (PESQ) [15], segmental SNRs and LLR. The enhanced speech components are estimated using (5.7). For loudness subtraction, the value of a is selected according to Figure 5.1. For loudness over-subtraction, a in Figure 5.1 is multiplied by a scaling factor as in Figure 5.5. The phase of the FFT of noisy speech is maintained for the estimated speech vector. The Inverse FFT is applied to derive the enhanced speech in the corresponding frame. Finally, the frame-by-frame speech signals is passed through a parallel to serial converter and the enhanced speech signal is reconstructed.

The noisy speech signal is enhanced with the spectral subtraction and over-subtraction, loudness subtraction (both the algorithm in [93] and the proposed algorithm) and over-subtraction. The quality of the enhanced signal is summarized in Tables 5.7 and 5.8. The subtraction factor a of the spectral over-subtraction is given as in Figure 5.3. The subtraction factor of these approaches in loudness domain is shown in Figure 5.6. The approach in [93] is a loudness subtraction algorithm with $a = 1$ for all the SNRs. We also include the adaptive-α algorithm proposed in the previous chapter for comparison.

Female Speech 1

Input SNR	Input Signal PESQ	PESQ score improvements					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.36	+0.13	+0.26	+0.17	+0.28	+0.40	+0.39
5	1.64	+0.17	+0.41	+0.21	+0.41	+0.50	+0.46
10	1.93	+0.19	+0.42	+0.25	+0.43	+0.54	+0.52
15	2.32	+0.21	+0.32	+0.26	+0.42	+0.50	+0.51
20	2.67	+0.18	+0.21	+0.24	+0.42	+0.51	+0.51

Female Speech 2

Input SNR	Input Signal PESQ	PESQ score improvements					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.60	+0.09	+0.29	+0.14	+0.20	+0.33	+0.20
5	1.87	+0.14	+0.21	+0.19	+0.17	+0.30	+0.30
10	2.17	+0.13	+0.16	+0.18	+0.23	+0.32	+0.31
15	2.44	+0.16	+0.14	+0.19	+0.28	+0.37	+0.34
20	2.73	+0.17	+0.18	+0.20	+0.35	+0.39	+0.38

Female Speech 3

Input SNR	Input Signal PESQ	PESQ score improvements					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.25	+0.18	+0.44	+0.24	+0.36	+0.53	+0.57
5	1.60	+0.21	+0.51	+0.27	+0.51	+0.61	+0.55
10	1.96	+0.23	+0.43	+0.28	+0.44	+0.57	+0.55
15	2.33	+0.22	+0.35	+0.27	+0.46	+0.53	+0.54
20	2.70	+0.21	+0.22	+0.25	+0.46	+0.53	+0.52

Table 5.7: Comparison of PESQ score improvements for subtraction-type speech enhancement algorithms (Female speech)

Male Speech 1

Input SNR	Input Signal PESQ	PESQ score improvements					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.39	+0.08	+0.17	+0.10	+0.17	+0.27	+0.31
5	1.57	+0.13	+0.33	+0.16	+0.38	+0.43	+0.42
10	1.83	+0.18	+0.36	+0.23	+0.40	+0.46	+0.49
15	2.15	+0.18	+0.26	+0.24	+0.36	+0.42	+0.41
20	2.49	+0.16	+0.05	+0.21	+0.25	+0.32	+0.31

Male Speech 2

Input SNR	Input Signal PESQ	PESQ score improvements					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.63	+0.12	+0.23	+0.16	+0.23	+0.35	+0.29
5	1.95	+0.21	+0.43	+0.27	+0.37	+0.53	+0.47
10	2.29	+0.25	+0.39	+0.30	+0.46	+0.55	+0.50
15	2.68	+0.26	+0.28	+0.32	+0.44	+0.50	+0.44
20	3.11	+0.19	+0.08	+0.22	+0.25	+0.33	+0.31

Male Speech 3

Input SNR	Input Signal PESQ	PESQ score improvements					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.35	+0.22	+0.49	+0.27	+0.47	+0.60	+0.53
5	1.70	+0.22	+0.52	+0.27	+0.52	+0.62	+0.63
10	2.06	+0.24	+0.50	+0.30	+0.54	+0.64	+0.61
15	2.47	+0.26	+0.40	+0.33	+0.52	+0.58	+0.55
20	2.84	+0.20	+0.26	+0.25	+0.42	+0.49	+0.51

Table 5.8: Comparison of PESQ score improvements for subtraction-type speech enhancement algorithms (Male speech)

Female Speech 1

Input SNR	Input Segmental SNR	Segmental SNR					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	-6.56	-3.14	1.48	-2.03	1.32	2.61	2.36
5	-1.56	1.43	2.58	2.43	4.52	5.31	5.36
10	3.44	5.81	4.05	6.69	7.76	8.32	8.23
15	8.44	10.18	5.78	10.90	11.13	11.52	11.49
20	13.44	14.61	7.58	15.10	14.74	14.88	14.94

Female Speech 2

Input SNR	Input Segmental SNR	Segmental SNR					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	-11.78	-8.26	-0.97	-7.12	-3.23	-1.47	-1.83
5	-6.78	-3.77	0.77	-2.74	0.28	1.67	1.46
10	-1.78	0.81	2.37	1.73	3.80	4.79	4.58
15	3.22	5.31	4.00	6.08	7.13	7.81	7.74
20	8.22	9.70	5.57	10.26	10.63	11.09	11.19

Female Speech 3

Input SNR	Input Segmental SNR	Segmental SNR					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	-6.75	-3.49	0.65	-2.40	0.47	1.70	1.55
5	-1.75	1.08	2.38	2.05	4.08	5.05	4.72
10	3.25	5.60	3.90	6.43	7.60	8.15	8.23
15	8.25	10.07	5.66	10.74	11.13	11.50	11.51
20	13.25	14.42	7.59	14.87	14.73	14.99	15.03

Table 5.9: Comparison of the Segmental SNR (in dB) of subtraction-type speech enhancement algorithms (Female speech)

Male Speech 1

Input SNR	Input Segmental SNR	Segmental SNR					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	-6.30	-2.88	1.29	-1.19	1.52	2.69	2.55
5	-1.30	1.76	2.84	2.76	5.21	6.03	6.04
10	3.70	6.32	4.38	7.19	8.70	9.32	9.24
15	8.70	10.62	6.19	11.37	12.00	12.46	12.35
20	13.70	15.05	8.06	15.59	15.42	15.64	15.65

Male Speech 2

Input SNR	Input Segmental SNR	Segmental SNR					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	-5.94	-2.84	0.98	-1.80	0.43	1.70	1.56
5	-0.94	1.50	2.43	2.41	3.88	4.74	4.81
10	4.06	6.03	4.15	6.81	7.46	8.01	7.88
15	9.06	10.49	5.87	11.08	11.01	11.39	11.31
20	14.06	14.97	7.82	15.37	14.99	15.16	15.17

Male Speech 3

Input SNR	Input Segmental SNR	Segmental SNR					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	-7.87	-4.63	0.58	-3.53	-0.41	0.98	0.90
5	-2.87	0.01	2.05	1.01	2.98	3.95	3.96
10	2.13	4.30	3.46	5.15	6.26	6.94	6.95
15	7.13	8.73	5.16	9.65	9.63	10.12	10.14
20	12.13	13.21	6.92	13.67	13.58	13.79	13.77

Table 5.10: Comparison of the Segmental SNR (in dB) of subtraction-type speech enhancement algorithms (Male speech)

Female Speech 1

Input SNR	Input LLR	Log-Likelihood Ratio (LLR)					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	2.02	1.81	1.60	1.76	1.85	1.62	1.55
5	1.65	1.40	1.19	1.35	1.43	1.21	1.25
10	1.24	1.00	0.91	0.95	1.04	0.91	0.90
15	0.90	0.74	0.71	0.68	0.74	0.65	0.65
20	0.61	0.48	0.54	0.45	0.50	0.46	0.48

Female Speech 2

Input SNR	Input LLR	Log-Likelihood Ratio (LLR)					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.59	1.52	1.45	1.47	1.74	1.52	1.47
5	1.35	1.22	1.15	1.16	1.42	1.21	1.19
10	1.03	0.91	0.91	0.87	1.21	0.99	0.92
15	0.80	0.72	0.83	0.68	0.97	0.84	0.79
20	0.57	0.50	0.62	0.47	0.65	0.57	0.53

Female Speech 3

Input SNR	Input LLR	Log-Likelihood Ratio (LLR)					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	2.09	1.88	1.57	1.81	1.69	1.56	1.50
5	1.71	1.43	1.16	1.37	1.26	1.12	1.14
10	1.31	1.08	0.99	1.01	1.12	1.03	0.88
15	0.93	0.75	0.71	0.70	0.76	0.63	0.66
20	0.67	0.54	0.61	0.50	0.55	0.48	0.49

Table 5.11: Comparison of the Log-Likelihood Ratio of subtraction-type speech enhancement algorithms (Female speech)

Male Speech 1

Input SNR	Input LLR	Log-Likelihood Ratio (LLR)					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	2.49	2.19	1.81	2.13	1.90	1.74	1.63
5	2.03	1.74	1.41	1.67	1.56	1.41	1.33
10	1.56	1.29	1.07	1.21	1.22	1.05	1.06
15	1.19	0.99	0.91	0.93	0.99	0.89	0.82
20	0.83	0.69	0.70	0.64	0.67	0.61	0.60

Male Speech 2

Input SNR	Input LLR	Log-Likelihood Ratio (LLR)					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.39	1.19	1.06	1.16	1.17	1.06	1.14
5	1.07	0.92	0.85	0.88	1.05	0.86	0.83
10	0.82	0.66	0.67	0.63	0.82	0.70	0.65
15	0.53	0.44	0.56	0.40	0.54	0.49	0.47
20	0.35	0.31	0.45	0.28	0.36	0.34	0.35

Male Speech 3

Input SNR	Input LLR	Log-Likelihood Ratio (LLR)					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.67	1.53	1.45	1.49	1.69	1.48	1.48
5	1.39	1.22	1.13	1.18	1.35	1.15	1.13
10	1.06	0.91	0.89	0.85	1.02	0.88	0.86
15	0.78	0.65	0.70	0.61	0.75	0.64	0.68
20	0.55	0.49	0.58	0.46	0.55	0.48	0.44

Table 5.12: Comparison of the Log-Likelihood Ratio of subtraction-type speech enhancement algorithms (Male speech)

Female Speech 1

Input SNR	Input $M_{overall}$	$M_{overall}$					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.65	1.87	2.08	1.92	1.97	2.18	2.19
5	2.07	2.33	2.64	2.39	2.51	2.70	2.64
10	2.51	2.79	3.02	2.86	2.96	3.12	3.10
15	3.00	3.25	3.36	3.32	3.42	3.53	3.54
20	3.43	3.64	3.64	3.71	3.83	3.92	3.91

Female Speech 2

Input SNR	Input $M_{overall}$	$M_{overall}$					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	2.07	2.18	2.37	2.24	2.15	2.37	2.29
5	2.41	2.59	2.68	2.66	2.51	2.72	2.73
10	2.81	2.98	3.00	3.04	2.91	3.09	3.12
15	3.15	3.32	3.25	3.36	3.29	3.43	3.43
20	3.50	3.68	3.62	3.71	3.74	3.81	3.83

Female Speech 3

Input SNR	Input $M_{overall}$	$M_{overall}$					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.53	1.78	2.15	1.87	2.02	2.23	2.29
5	2.01	2.32	2.70	2.39	2.65	2.80	2.74
10	2.50	2.80	3.01	2.88	2.95	3.10	3.16
15	2.99	3.26	3.39	3.33	3.45	3.57	3.57
20	3.42	3.66	3.63	3.71	3.85	3.95	3.85

Table 5.13: Comparison of the composite measure $M_{overall}$ of subtraction-type speech enhancement algorithms (Female speech)

Male Speech 1

Input SNR	Input $M_{overall}$	$M_{overall}$					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.44	1.66	1.92	1.70	1.88	2.04	2.12
5	1.82	2.07	2.40	2.13	2.37	2.48	2.51
10	2.27	2.55	2.81	2.63	2.76	2.90	2.91
15	2.72	2.96	3.07	3.04	3.11	3.21	3.23
20	3.17	3.37	3.28	3.44	3.46	3.54	3.54

Male Speech 2

Input SNR	Input $M_{overall}$	$M_{overall}$					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	2.19	2.39	2.55	2.44	2.49	2.65	2.56
5	2.62	2.86	3.07	2.93	2.92	3.15	3.12
10	3.02	3.30	3.41	3.36	3.39	3.52	3.51
15	3.48	3.74	3.69	3.80	3.83	3.90	3.87
20	3.92	4.09	3.93	4.13	4.11	4.19	4.17

Male Speech 3

Input SNR	Input $M_{overall}$	$M_{overall}$					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.83	2.07	2.33	2.14	2.19	2.41	2.35
5	2.25	2.52	2.80	2.58	2.69	2.87	2.89
10	2.71	2.98	3.19	3.06	3.16	3.32	3.30
15	3.18	3.46	3.55	3.54	3.62	3.72	3.68
20	3.60	3.79	3.79	3.86	3.94	4.03	4.07

Table 5.14: Comparison of the composite measure $M_{overall}$ of subtraction-type speech enhancement algorithms (Male speech)

Female Speech 1

Input SNR	Input M_{back}	M_{back}					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.87	2.15	2.50	2.24	2.50	2.64	2.61
5	2.32	2.59	2.77	2.67	2.90	3.00	2.98
10	2.77	3.01	3.01	3.10	3.25	3.34	3.32
15	3.27	3.48	3.26	3.55	3.64	3.71	3.71
20	3.76	3.92	3.48	3.98	4.04	4.09	4.10

Female Speech 2

Input SNR	Input M_{back}	M_{back}					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.66	1.92	2.48	2.01	2.29	2.46	2.38
5	2.10	2.36	2.68	2.45	2.63	2.78	2.76
10	2.56	2.78	2.90	2.87	3.02	3.13	3.11
15	3.00	3.21	3.12	3.27	3.38	3.47	3.45
20	3.46	3.63	3.38	3.68	3.78	3.82	3.83

Female Speech 3

Input SNR	Input M_{back}	M_{back}					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.81	2.10	2.48	2.20	2.43	2.59	2.60
5	2.29	2.57	2.79	2.66	2.90	3.01	2.96
10	2.78	3.03	3.02	3.11	3.26	3.36	3.35
15	3.27	3.49	3.27	3.55	3.67	3.73	3.73
20	3.76	3.93	3.51	3.98	4.07	4.12	4.07

Table 5.15: Comparison of the composite measure M_{back} of subtraction-type speech enhancement algorithms (Female speech)

Male Speech 1

Input SNR	Input M_{back}	M_{back}					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.90	2.16	2.46	2.27	2.48	2.60	2.60
5	2.30	2.56	2.72	2.63	2.89	2.97	2.96
10	2.74	2.99	2.96	3.07	3.25	3.32	3.32
15	3.21	3.42	3.18	3.49	3.59	3.65	3.63
20	3.69	3.85	3.36	3.91	3.92	3.96	3.95

Male Speech 2

Input SNR	Input M_{back}	M_{back}					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	2.04	2.29	2.58	2.38	2.55	2.69	2.65
5	2.51	2.76	2.92	2.85	2.99	3.12	3.09
10	2.98	3.23	3.18	3.30	3.42	3.50	3.46
15	3.49	3.70	3.42	3.77	3.82	3.87	3.84
20	4.01	4.15	3.65	4.19	4.18	4.23	4.22

Male Speech 3

Input SNR	Input M_{back}	M_{back}					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.78	2.09	2.55	2.19	2.48	2.63	2.59
5	2.27	2.55	2.82	2.64	2.88	2.99	3.00
10	2.75	3.00	3.08	3.09	3.27	3.36	3.35
15	3.26	3.49	3.33	3.58	3.67	3.73	3.72
20	3.75	3.92	3.55	3.97	4.05	4.09	4.10

Table 5.16: Comparison of the composite measure M_{back} of subtraction-type speech enhancement algorithms (Male speech)

Female Speech 1

Input SNR	Input M_{sig}	M_{sig}					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.83	2.13	2.42	2.20	2.18	2.49	2.54
5	2.38	2.74	3.10	2.82	2.86	3.14	3.07
10	2.98	3.34	3.57	3.43	3.45	3.65	3.63
15	3.57	3.86	3.95	3.95	3.98	4.12	4.13
20	4.07	4.32	4.27	4.38	4.44	4.54	4.52

Female Speech 2

Input SNR	Input M_{sig}	M_{sig}					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	2.42	2.55	2.74	2.63	2.39	2.69	2.67
5	2.83	3.05	3.16	3.14	2.86	3.16	3.18
10	3.34	3.54	3.56	3.61	3.30	3.58	3.64
15	3.74	3.92	3.79	3.98	3.74	3.92	3.96
20	4.15	4.33	4.21	4.38	4.28	4.39	4.42

Female Speech 3

Input SNR	Input M_{sig}	M_{sig}					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.70	2.02	2.50	2.13	2.32	2.56	2.65
5	2.30	2.71	3.17	2.81	3.07	3.27	3.22
10	2.93	3.30	3.52	3.40	3.38	3.56	3.70
15	3.54	3.86	3.98	3.94	3.99	4.17	4.14
20	4.03	4.29	4.23	4.36	4.43	4.55	4.47

Table 5.17: Comparison of the composite measure M_{sig} of subtraction-type speech enhancement algorithms (Female speech)

Male Speech 1

Input SNR	Input M_{sig}	M_{sig}					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.37	1.73	2.17	1.80	2.08	2.30	2.44
5	1.95	2.33	2.79	2.42	2.66	2.85	2.92
10	2.59	2.98	3.31	3.09	3.18	3.39	3.40
15	3.16	3.48	3.61	3.58	3.59	3.73	3.79
20	3.74	3.98	3.90	4.06	4.06	4.16	4.16

Male Speech 2

Input SNR	Input M_{sig}	M_{sig}					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	2.65	2.92	3.12	2.98	3.01	3.20	3.07
5	3.17	3.45	3.65	3.53	3.41	3.70	3.70
10	3.63	3.95	4.02	4.01	3.91	4.09	4.11
15	4.16	4.41	4.30	4.49	4.42	4.51	4.49
20	4.61	4.76	4.55	4.81	4.75	4.82	4.80

Male Speech 3

Input SNR	Input M_{sig}	M_{sig}					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	2.19	2.47	2.71	2.54	2.45	2.75	2.70
5	2.69	3.00	3.27	3.07	3.04	3.31	3.34
10	3.24	3.54	3.72	3.64	3.61	3.82	3.82
15	3.78	4.07	4.10	4.15	4.12	4.27	4.21
20	4.24	4.42	4.37	4.48	4.49	4.61	4.66

Table 5.18: Comparison of the composite measure M_{sig} of subtraction-type speech enhancement algorithms (Male speech)

As the loudness model is more suited for simulating human hearing systems and evaluating the speech quality, the loudness subtraction and over-subtraction outperform the corresponding methods in the spectral domain in the PESQ scores. The loudness subtraction provides 0.02-0.06 improvement over the spectral subtraction. The loudness over-subtraction provides 0.04-0.17 improvement over the spectral over-subtraction. Compared to the previous loudness subtraction approach in [93], our proposed loudness over-subtraction method provides consistent improvement of 0.04-0.31. The adaptive-α approach provides 0-0.21 improvement over the spectral over-subtraction approach.

Again the quality of the enhancement speech is measured using the segmental SNR. As seen in Tables 5.9 and 5.10, the loudness over-subtraction has better segmental SNR results than the other approaches in all but one cases (female speech 2 at 0 dB, approach in [93]). Overall, the proposed loudness subtraction outperforms spectral subtraction by 0.40-1.69 dB. The proposed loudness over-subtraction outperforms spectral over-subtraction by 0.14-1.76 dB. Especially when the input SNR is low (0 dB), the difference exceeds 1dB for both subtraction and over-subtraction. Again, the adaptive-α approach is better than the spectral over-subtraction by 0.18-1.40 dB. The loudness over-subtraction performs slightly better or comparable to adaptive-α approach. The differences are below 0.25 dB for the tested speeches.

As the approach in [93] is also applied in the loudness domain, it can provide excellent segmental SNR improvement when the input SNR is low. When the input SNR is high, the approach over-subtracts considerably and can even lead to segmental SNR losses. On the other hand, the proposed loudness subtraction and over-subtraction can provide consistent improvement over the unprocessed noisy signal. The loudness over-subtraction outperforms the approach in [93] in all but one case, even when the SNR is low.

Overall, we can conclude that the loudness subtraction and over-subtraction can provide better noise removal than the corresponding approaches in the spectral domain. The proposed loudness over-subtraction approach also outperforms the previously proposed loudness subtraction approach in [93]. Simulation results have shown the proposed loudness-subtraction approach to be a better subtraction-type speech enhancement algorithm with regards to noise removal.

To better assess the quality, we now consider the performance based on log likelihood ratio. In Table 5.11 and 5.12, the log likelihood ratio of the noisy and enhanced speech signals are compared. Usually the speech enhancement algorithms have to compromise between the noise removal and signal distortion. For example, spectral over-subtraction provides better noise removal and PESQ scores, but it does increase the signal distortion measured by the Log-likelihood ratio.

The loudness subtraction approach provides better (i.e. smaller) LLRs compared to the spectral subtraction when the input SNR is high and comparable LLRs when the input SNR is low. At the same time, the loudness subtraction leads to better noise removal measured by segmental SNR and better overall quality measured by PESQ.

For the loudness over-subtraction and adaptive-α, it outperforms the spectral over-subtraction in all categories. It provides better noise removal, less distortion and improved overall quality over spectral over-subtraction. They can provide similar or even less distortion than the spectral subtraction, while significantly improving the segmental SNR and PESQ scores.

The results confirm that using the loudness over-subtraction of noise ensures that as much noise as possible is removed while not resulting in increased distortion of the signal. The loudness over-subtraction has been shown to provide the best noise removal and overall

quality. It also has similar or less distortion compared to spectral subtraction approach.

In the Tables 5.13 to 5.18, a set of composite quality measures are used to evaluate the performance:

$$M_{overall} = 1.594 - 0.512 \cdot LLR + 0.805 \cdot PESQ \tag{5.13}$$

$$M_{back} = 1.634 + 0.063 \cdot segSNR + 0.478 \cdot PESQ \tag{5.14}$$

$$M_{sig} = 3.093 - 1.029 \cdot LLR + 0.603 \cdot PESQ \tag{5.15}$$

where a higher value implies better quality. $M_{overall}$, M_{back} and M_{sig} are the evaluations of the overall quality, the background noise and the speech signal, respectively. The results confirms that the proposed loudness over-subtraction and adaptive-α algorithms do improve the overall quality, the background noise reduction and distortion over the comparable algorithms. Next, we evaluate the performance of our algorithms for non-white noises.

Tables 5.19 to 5.22 shows the performances of subtraction-type speech enhancement algorithms under non-white noises. Two types of noises are used, the F16 noise and the babble noise [16].

Preliminary results shows that the proposed loudness over-subtraction and adaptive-α algorithms still outperform the spectral over-subtraction methods. The PESQ improvements, Segmental SNR and LLR all improve with the proposed loudness over-subtraction and adaptive-α algorithms. The results from the two proposed algorithms are comparable.

For the F16 noise, the improvement is significant for PESQ and segmental SNR. For the Babble noise, the improvement is not as significant comparing to the noisy speech. This is due to the fact that additive babble noise has frequency characteristics similar to the speech itself, which makes it more difficult to separate them one from the other.

Input SNR	Input Signal PESQ	PESQ score improvements					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.47	+0.11	+0.26	+0.19	+0.09	+0.22	+0.21
5	1.90	+0.10	+0.20	+0.16	+0.10	+0.18	+0.17
10	2.30	+0.10	+0.13	+0.15	+0.07	+0.16	+0.14
15	2.70	+0.11	+0.03	+0.17	+0.10	+0.15	+0.14
20	3.06	+0.14	-0.05	+0.17	+0.15	+0.21	+0.21

Input SNR	Input Seg-mental SNR	Segmental SNR					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	-6.58	-3.94	0.00	-2.97	-1.42	-0.27	-0.40
5	-1.58	0.54	1.47	1.35	2.01	2.87	2.78
10	3.42	5.00	3.06	5.65	5.74	6.27	6.24
15	8.42	9.51	4.80	9.99	9.73	9.99	10.00
20	13.42	14.13	6.69	14.40	14.04	14.10	14.16

Input SNR	Input LLR	Log-Likelihood Ratio (LLR)					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	0.95	0.96	1.15	0.94	1.33	1.19	1.19
5	0.74	0.72	0.89	0.70	0.96	0.87	0.87
10	0.53	0.51	0.66	0.50	0.65	0.61	0.61
15	0.36	0.35	0.52	0.34	0.47	0.43	0.43
20	0.23	0.23	0.40	0.22	0.33	0.31	0.31

Table 5.19: Comparison of the objective quality measures of subtraction-type speech enhancement algorithms (Female speech 3, babble noise)

Input SNR	Input Signal PESQ	PESQ score improvements					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.66	+0.08	+0.19	+0.14	+0.11	+0.20	+0.19
5	2.02	+0.10	+0.22	+0.16	+0.14	+0.22	+0.21
10	2.37	+0.13	+0.26	+0.17	+0.20	+0.29	+0.28
15	2.74	+0.16	+0.21	+0.22	+0.23	+0.29	+0.28
20	3.13	+0.13	+0.09	+0.17	+0.21	+0.27	+0.27

Input SNR	Input Seg-mental SNR	Segmental SNR					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	-7.84	-5.22	-0.06	-4.27	-2.09	-0.80	-0.98
5	-2.84	-0.70	1.52	0.14	1.45	2.42	2.29
10	2.16	3.82	3.13	4.52	5.23	5.88	5.80
15	7.16	8.39	4.83	8.94	9.23	9.58	9.57
20	12.16	13.06	6.66	13.39	13.50	13.58	13.61

Input SNR	Input LLR	Log-Likelihood Ratio (LLR)					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.07	1.03	1.13	1.01	1.34	1.21	1.17
5	0.82	0.77	0.86	0.74	1.04	0.91	0.87
10	0.60	0.55	0.66	0.53	0.76	0.66	0.65
15	0.42	0.41	0.53	0.37	0.53	0.47	0.47
20	0.28	0.26	0.44	0.25	0.41	0.34	0.35

Table 5.20: Comparison of the objective quality measures of subtraction-type speech enhancement algorithms (Male speech 3, babble noise)

Input SNR	Input Signal PESQ	PESQ score improvements					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.45	+0.17	+0.42	+0.25	+0.35	+0.47	+0.46
5	1.87	+0.19	+0.37	+0.25	+0.32	+0.43	+0.41
10	2.28	+0.17	+0.25	+0.23	+0.30	+0.36	+0.35
15	2.67	+0.16	+0.14	+0.22	+0.27	+0.33	+0.32
20	3.03	+0.16	+0.06	+0.16	+0.27	+0.31	+0.31

Input SNR	Input Seg-mental SNR	Segmental SNR					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	-5.94	-2.94	0.69	-1.97	0.42	1.53	1.37
5	-0.94	1.49	2.10	2.40	3.75	4.53	4.53
10	4.06	5.92	3.66	6.65	7.20	7.72	7.67
15	9.06	10.39	5.39	10.91	10.94	11.21	11.22
20	14.06	14.90	7.29	15.25	14.98	15.03	15.09

Input SNR	Input LLR	Log-Likelihood Ratio (LLR)					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.24	1.14	1.10	1.08	1.36	1.14	1.11
5	0.97	0.85	0.85	0.81	1.10	0.89	0.87
10	0.70	0.61	0.68	0.57	0.83	0.67	0.66
15	0.48	0.41	0.53	0.40	0.56	0.46	0.46
20	0.30	0.27	0.41	0.26	0.38	0.31	0.31

Table 5.21: Comparison of the objective quality measures of subtraction-type speech enhancement algorithms (Female speech 3, F16 noise)

Input SNR	Input Signal PESQ	PESQ score improvements					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.60	+0.17	+0.39	+0.24	+0.31	+0.48	+0.46
5	1.99	+0.17	+0.40	+0.25	+0.34	+0.46	+0.44
10	2.37	+0.18	+0.35	+0.25	+0.32	+0.41	+0.40
15	2.75	+0.17	+0.22	+0.22	+0.27	+0.35	+0.34
20	3.12	+0.15	+0.11	+0.20	+0.29	+0.34	+0.34

Input SNR	Input Seg-mental SNR	Segmental SNR					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	-7.10	-3.96	0.52	-2.95	-0.11	1.06	0.89
5	-2.10	0.48	2.01	1.38	3.10	3.98	3.88
10	2.90	4.89	3.57	5.66	6.47	7.08	7.02
15	7.90	9.39	5.27	9.94	10.14	10.50	10.49
20	12.90	13.92	7.14	14.28	14.20	14.30	14.36

Input SNR	Input LLR	Log-Likelihood Ratio (LLR)					
		Spectral subtraction [80]	Approach in [93]	Loudness subtraction	Spectral over-subtraction [25]	Loudness over-subtraction	adaptive α subtraction
0	1.14	1.10	1.20	1.18	1.41	1.28	1.25
5	0.89	0.70	0.83	0.73	1.03	0.90	0.87
10	0.67	0.55	0.66	0.53	0.72	0.63	0.62
15	0.49	0.41	0.52	0.37	0.55	0.47	0.45
20	0.33	0.26	0.44	0.25	0.41	0.32	0.31

Table 5.22: Comparison of the objective quality measures of subtraction-type speech enhancement algorithms (Male speech 3, F16 noise)

5.5.2 Comparison of Subtraction Type Algorithms

In this section, we rewrite the adaptive-α and loudness over-subtraction equations in the form $\hat{X} = GY$, where the enhancement is achieved by multiplying the noisy speech by a gain G. The gain function G is given by:

$$G = \frac{\hat{X}}{Y} = \left(1 - a\left(\frac{N^2}{Y^2}\right)^{\alpha}\right)^{\frac{1}{2\alpha}} = \left(1 - a\left(\frac{1}{\gamma}\right)^{\alpha}\right)^{\frac{1}{2\alpha}} \tag{5.16}$$

where $\gamma = \frac{Y^2}{N^2}$ is the instantaneous noisy signal-to-noise ratio (NSNR). G is thus expressed as a function of the *a priori* SNR ξ and the instantaneous SNR $\gamma - 1$. ξ is used to calculate the corresponding α in the adaptive-α approach and a in the loudness over-subtraction approach. We also compare the results with the spectral over-subtraction approach, which corresponds to fixed $\alpha = 1$ and adaptive a with ξ.

From the Figure 5.7, we can clearly see that the spectral over-subtraction approach did very little to enhance the noisy speech signal when the instantaneous SNR is high and complete eliminate the signal vice versa. On the other hand, our proposed approaches reduce the noise when the instantaneous SNR is high and preserve some of the signal when the instantaneous SNR is low.

As expected, the proposed adaptive-α and loudness over-subtraction has similar gain functions. The two algorithms use different setups but achieve similar results.

5.6 An Improved NSNR Estimation

The NSNR is estimated using the estimated noisy signal variance and the estimated noise variance. Usually, the estimation of the noise variance is obtained using longer time constants. However, due to the non-stationary nature of speech, the time constant used for noisy speech variance estimation is relatively short.

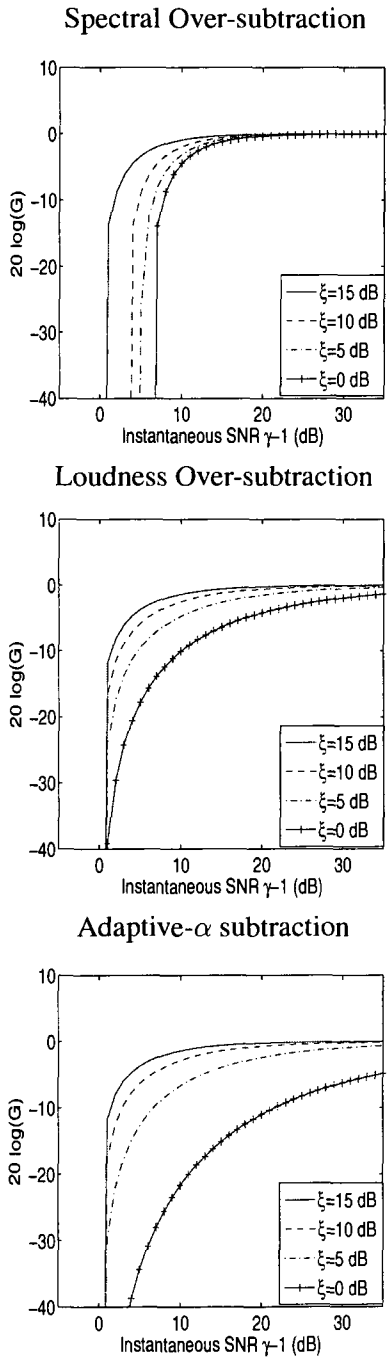

Figure 5.7: The gain functions of the subtraction type approaches

The regular update of the noisy speech signal variance is:

$$\sigma_y^2(n) = \beta\sigma_y^2(n) + (1-\beta)\left|y^2(n)\right| \tag{5.17}$$

and the β is chosen to have a time constant of 10 ms. In this case $\beta = 0.4493$.

For our proposed loudness subtraction method, the value chosen for a is more sensitive to the NSNR when the NSNR is low (as in Figure 5.1). Also when the NSNR is low, the scaling factor is large, which brings more variation to the determination of a in the loudness over-subtraction method. Thus, we need a more accurate estimation of the NSNR in this case.

One such method is to estimate the noisy signal energy with a larger time constant when the NSNR is low. As the overall NSNR will change relatively slower from frame to frame when the NSNR is low, the larger time constant is acceptable without dramatically influencing the tracking ability of NSNR estimation. Then the update for the noisy signal energy can be made adaptive for different NSNRs. When the NSNR is low, a larger time constant is chosen (larger β) to reduce the error of NSNR estimation. When the NSNR is high, the time constant will be chosen as stated before.

A simple choice is used as shown in Figure 5.8 where β is chosen based on the value of the NSNR estimation from the previous frame. Results are obtained for a fixed β value and then repeated for the Adaptive-β proposed above (the scaling multiplier type loudness over-subtraction with [4,1]@[0,20] dB). The results are shown in Tables 5.23 and 5.24.

From these tables, we can see that adaptive-β SNR estimation can improve the PESQ scores when the SNR is 0 dB. The improvements are more than 0.1 for 5 out of 6 cases. For the other cases and SNRs, the PESQ scores are similar or slightly worse.

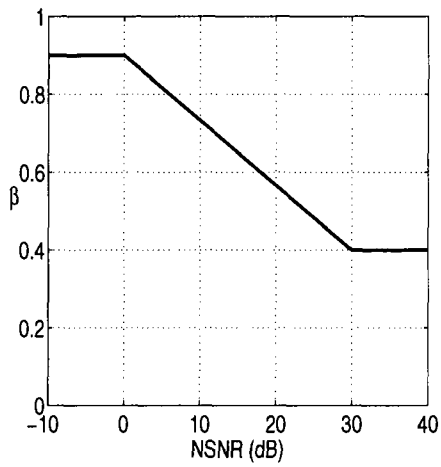

Figure 5.8: The chosen value of β with the NSNR

SNR (dB)	PESQ score improvements					
	Speech 1		Speech 2		Speech 3	
	Fixed β	Adaptive-β	Fixed β	Adaptive-β	Fixed β	Adaptive-β
0	+0.40	+0.58	+0.33	+0.34	+0.53	+ 0.68
5	+0.50	+0.56	+0.30	+0.25	+0.61	+0.68
10	+0.54	+0.54	+0.32	+0.27	+0.56	+0.58
15	+0.50	+0.44	+0.37	+0.36	+0.53	+0.50
20	+0.51	+0.49	+0.39	+0.39	+0.53	+0.50

Table 5.23: Comparison of PESQ score improvements with fixed β and adaptive-β (Female speech)

SNR (dB)	PESQ score improvements					
	Speech 1		Speech 2		Speech 3	
	Fixed β	Adaptive-β	Fixed β	Adaptive-β	Fixed β	Adaptive-β
0	+0.27	+0.38	+0.35	+0.49	+0.60	+ 0.79
5	+0.43	+0.40	+0.53	+0.57	+0.62	+0.69
10	+0.46	+0.42	+0.54	+0.53	+0.64	+0.64
15	+0.42	+0.35	+0.50	+0.48	+0.58	+0.53
20	+0.32	+0.24	+0.33	+0.32	+0.49	+0.46

Table 5.24: Comparison of PESQ score improvements with fixed β and adaptive-β (Male speech)

5.7 Conclusions

In this chapter, we presented speech enhancement approaches based on loudness subtraction and over-subtraction. Several speech enhancement algorithms based on spectral subtraction and over-subtraction have been reviewed. The fact that the loudness model was shown to be better matched to the human hearing system has motivated the research of the loudness subtraction for speech enhancement [93]. This loudness subtraction approach subtracts the loudness of the estimated noise directly from the loudness of the noisy speech signal. Based on simulations, the enhanced signal has a noticeable distortion when the SNR is high.

On the average, the sum of the noise loudness and clean speech loudness is not equal to the noisy speech loudness. This is especially true when the SNR is high. In this chapter, the relationship between the loudness of these three signals was developed with the assumption of the Gaussian model for noise and Laplacian model for speech and that both are stationary. We propose a new way to determine the subtraction factor a so that the three loudness parameters can be balanced statistically.

Based on this determination of the subtraction factor, a new loudness subtraction method is proposed. The subtraction factor is selected adaptively for each frame based on the NSNR of the speech signal assuming Laplacian distribution of the speech signal and Gaussian distribution for the additive noise. This new approach performs consistently for all input SNRs.

An over-subtraction approach has been introduced in spectral subtraction [25]. The noise is usually assumed to be Gaussian. When we subtract the average spectrum (or loudness) of the noise, some large noise samples will still remain in the enhanced speech, which can be very disturbing. The over-subtraction approach eliminates the large noise

samples at low SNRs, while keep the regular subtraction at high SNRs. Experimentally, this approach leads to better performance of speech enhancement.

In this chapter, we extended the spectral over-subtraction to the loudness over-subtraction approach. The results show that the chosen value of the multiplier may vary with different input SNRs. Like the spectral subtraction, a larger scaling factor should be assigned when the SNR is low and vice versa. This leads an adaptive loudness over-subtraction, which was shown to provide consistent improvements over a broad range of input SNRs.

The proposed loudness subtraction is compared to an existing direct loudness subtraction models, which is equivalent to a loudness over-subtraction model with a larger scaling factor when the SNR is high and vice versa. This directly leads to good noise removal at low SNRs, but unsatisfactory performance at high SNRs. Our proposed over-subtraction better compensates for the noise and performs consistently over all SNRs.

The proposed subtraction-type speech enhancement algorithms are compared to several existing ones using three objective measures: PESQ, segmental SNR and LLR. The results have shown that the loudness subtraction and over-subtraction outperform the corresponding spectral domain methods in the PESQ and segmental SNR. The loudness subtraction and over-subtraction can lead to less or comparable distortion measured by LLR. Overall, the loudness over-subtraction has been shown to be the best method for subtraction-type speech enhancement. It removes the noise considerably without bringing further distortion to the enhanced signal. All the evaluations confirm that the research in the loudness domain can provide a better environment for speech enhancement rather than the spectral domain.

Also, as the subtraction factor depends on the SNR or NSNR estimation of the noisy speech frames, the estimation of the SNR or NSNR is critical. Especially when the SNR is low, the subtraction factor or scaling factor can vary dramatically with small difference in

SNR. As the speech is non-stationary over longer periods, the time constant chosen for the estimation of the variance of noisy speech is usually comparable to the length of the frame. When the overall SNR is low, the SNR will change slower from frame to frame. This allows the use of a longer time constant when the SNR is low. Hence an improved version of NSNR estimation method was proposed to adapt the time constant to the SNRs. It was shown to achieve better performance when the SNR is low. Overall, these approaches in the loudness domain result in improved SNR as well as improved PESQ scores and less musical noise in the informal subjective listening.

Chapter 6

Speech Enhancement Based on Maximum Likelihood (ML) Estimation

6.1 Introduction

In the speech enhancement process, the speech and noise signals can be characterized with probability distribution functions (PDFs) and a set of parameters that define the PDFs. Then we can determine the estimated clean speech from the noisy speech signal. Bayesian parameter estimation [5] minimized the cost function that averaging over the conditional density function of the possible parameter (clean speech) given the observation sample (noisy speech). Minimum mean square error (MMSE) estimator uses the mean square error as the cost function and is applied in speech enhancement [36] [111] [114]. The main factors in the MMSE type algorithms are the designation of the cost function to be minimized and the statistical assumption of the variables.

For the MMSE estimator, the definition of the error function is critical. Different error

functions or criteria have been used: the short time spectral amplitude (STSA) [41], the log-STSA [42], β-order spectral amplitude [114] and speech magnitude-squared spectra [36].

In most previous research, the speech is assumed to have a Gaussian distribution [41]. For the short speech frames that are used by the speech enhancement algorithms (usually around 10-40 ms), the Laplacian model has been shown to be superior to the Gaussian model [47].

A Laplacian-based MMSE speech enhancement algorithm has been proposed in [30]. The MMSE is applied to the magnitude spectrum of the speech signals in the DFT domain. The Laplacian distribution is assumed for the speech signals and Gaussian for the additive noise. The analytical solution of the MMSE estimation is given using the infinite summation of the hypergeometric functions. This approach provides significant improvement of quality over the Gaussian based MMSE approach. Even though the authors use finite approximations, the complexity of this algorithm is considerably higher than the previous Gaussian approach.

Since the MMSE algorithm is usually computationally complex, Maximum Likelihood (ML) estimation [48] [86] [118] has been used as a lower complexity alternative. In [77], the speech signal is modeled as a sum of sinusoids, the likelihood functions of which were maximized.

In this Chapter, we first review the MMSE-based speech enhancement algorithms. Then we will discuss the maximum likelihood-based speech enhancement, especially in the loudness domain. We proposed ML speech enhancement based on Gaussian and Laplacian speech models. Also performance of the algorithms will be compared to the MMSE algorithms.

6.2 Review of the MMSE-Based Speech Enhancement Algorithms

In [114], a generalized MMSE method is proposed. The cost function C is defined as:

$$C = \left(\hat{s}^{2\alpha} - s^{2\alpha}\right)^2, \tag{6.1}$$

where s is the FFT magnitude of the clean speech signal, $\hat{s}$ is the FFT amplitude of the enhanced speech and α is a parameter that can be adjust to allow the algorithm to minimize the error in different domains. For example, α=1 minimizes the error in the power spectral domain, α=0.5 minimizes the error in the spectral domain and α=0.27 minimizes the error in the loudness domain.

The MMSE solution for $\hat{s}^{2\alpha}$ is given by:

$$\begin{aligned}
\hat{s}^{2\alpha} &= E\left\{s^{2\alpha}|Y\right\} \\
&= \int s^{2\alpha} p(s|Y) ds \\
&= \int s^{2\alpha} \frac{p(s,Y)}{p(Y)} ds \\
&= \frac{\int s^{2\alpha} p(s,Y) ds}{p(Y)} \\
&= \frac{\int s^{2\alpha} p(Y|s) p(s) ds}{p(Y)} \\
&= \frac{\int_0^{\infty} \int_0^{2\pi} s^{2\alpha} p\left(Y|s,\theta\right) p\left(s,\theta\right) d\theta ds}{\int_0^{\infty} \int_0^{2\pi} p\left(Y|s,\theta\right) p\left(s,\theta\right) d\theta ds}
\end{aligned} \tag{6.2}$$

where Y is the FFT sample of the noisy speech signal and $p(s,Y)$ is the joint PDF. $p(Y|s)$ are the conditional probability functions of Y given s and $p(s|Y)$ are the conditional probability functions of s given Y. This above solution will be discussed for different statistical assumptions for the speech signals: Gaussian and Laplacian.

6.2.1 MMSE-Based Speech Enhancement Algorithm with Gaussian Speech Model

In this section, we propose a speech enhancement algorithm when the estimated speech is determined based on (6.2). The algorithm is presented for a Gaussian model for both speech and noise.

The Gaussian model assumes that in the frequency domain, the noise $N = ue^{j\varphi}$ and the clean speech signal $X = se^{j\theta}$ both following the Gaussian distribution with the variances of λ_N and λ_X, and $Y = X + N$. So that the terms in (6.2) can be written as:

$$p(Y|s,\theta) = \frac{1}{\pi\lambda_N} \exp\left\{-\frac{1}{\lambda_N}\left|Y - se^{j\theta}\right|^2\right\} \tag{6.3}$$

and

$$p(s,\theta) = \frac{s}{\pi\lambda_X} \exp\left\{-\frac{s^2}{\lambda_X}\right\}. \tag{6.4}$$

We will now show how to practically calculate the estimated speech given in (6.2). Using (6.3) and (6.4), the numerator term in (6.2) can be written as:

$$\begin{aligned}
&\int_0^\infty \int_0^{2\pi} s^{2\alpha} p(Y|s,\theta)\, p(s,\theta)\, d\theta ds \\
&= \int_0^\infty \int_0^{2\pi} s^{2\alpha} \frac{1}{\pi\lambda_N} \exp\left\{-\frac{1}{\lambda_N}\left|Y - se^{j\theta}\right|^2\right\} \frac{s}{\pi\lambda_X} \exp\left\{-\frac{s^2}{\lambda_X}\right\} d\theta ds \\
&= \int_0^\infty s^{2\alpha} \frac{s}{\pi^2\lambda_X\lambda_N} \exp\left\{-\frac{s^2}{\lambda_X}\right\} \int_0^{2\pi} \exp\left\{-\frac{1}{\lambda_N}\left|Y - se^{j\theta}\right|^2\right\} d\theta ds.
\end{aligned} \tag{6.5}$$

The integral term with reference to θ will be simplified first. Assume that $Y = ve^{j\gamma}$,

we obtain [2, eq.9.215.1]:

$$\begin{aligned}
&\int_0^{2\pi} \exp\left\{-\frac{1}{\lambda_N}\left|Y - se^{j\theta}\right|^2\right\} d\theta \\
&= \int_0^{2\pi} \exp\left\{-\frac{1}{\lambda_N}\left|v\cos\gamma + jv\sin\gamma - s\cos\theta - js\sin\theta\right|^2\right\} d\theta \\
&= \int_0^{2\pi} \exp\left\{-\frac{1}{\lambda_N}\left((v\cos\gamma - s\cos\theta)^2 + (v\sin\gamma - s\sin\theta)^2\right)\right\} d\theta \\
&= \int_0^{2\pi} \exp\left\{-\frac{1}{\lambda_N}\left(v^2 + s^2 - 2vs\left(\cos\gamma\cos\theta + \sin\gamma\sin\theta\right)\right)\right\} d\theta \\
&= \int_0^{2\pi} \exp\left\{-\frac{1}{\lambda_N}\left(v^2 + s^2\right)\right\} \exp\left\{\frac{1}{\lambda_N}\left(2vs\cdot\cos\left(\gamma - \theta\right)\right)\right\} d\theta \\
&= \exp\left\{-\frac{1}{\lambda_N}\left(v^2 + s^2\right)\right\} \int_0^{2\pi} \exp\left\{\frac{1}{\lambda_N}\left(2vs\cdot\cos\left(\theta\right)\right)\right\} d\theta \\
&= 2\pi \exp\left\{-\frac{1}{\lambda_N}\left(v^2 + s^2\right)\right\} I_0\left(\frac{2vs}{\lambda_N}\right),
\end{aligned} \tag{6.6}$$

where I_0 is the zero order modified Bessel function.

Now, (6.5) can be simplified [2, eq. 6.631.1, 9.210.1] as:

$$\begin{aligned}
&\int_0^{\infty}\int_0^{2\pi} s^{2\alpha} p\left(Y|s,\theta\right) p\left(s,\theta\right) d\theta ds \\
&= \int_0^{\infty} \frac{s^{2\alpha+1}}{\pi^2\lambda_N\lambda_X} \exp\left\{-\frac{s^2}{\lambda_X} - \frac{1}{\lambda_N}\left(v^2 + s^2\right)\right\} I_0\left(\frac{2vs}{\lambda_N}\right) ds \\
&= \frac{1}{\pi^2\lambda_N\lambda_X} \exp\left\{-\frac{1}{\lambda_N}v^2\right\} \int_0^{\infty} s^{2\alpha+1} \exp\left\{-\left(\frac{1}{\lambda_X} + \frac{1}{\lambda_N}\right)s^2\right\} J_0\left(j\frac{2v}{\lambda_N}s\right) ds \\
&= \frac{1}{\pi^2\lambda_N\lambda_X} \exp\left\{-\frac{1}{\lambda_N}v^2\right\} \frac{\Gamma\left(\frac{2\alpha+1}{2} + \frac{1}{2}\right)}{2\left(\frac{1}{\lambda_X} + \frac{1}{\lambda_N}\right)^{\frac{2\alpha+1}{2}+\frac{1}{2}}\Gamma\left(1\right)} \Phi\left(\frac{2\alpha+1}{2} + \frac{1}{2}; 1; \beta\right) \\
&= C\frac{\Gamma\left(\alpha+1\right)}{2\left(\frac{1}{\lambda_X} + \frac{1}{\lambda_N}\right)^{\alpha+1}} \Phi\left(\alpha+1; 1; \beta\right)
\end{aligned} \tag{6.7}$$

where J_0 is the zero order Bessel function, C is:

$$C = \frac{1}{\pi^2\lambda_N\lambda_X} \exp\left\{-\frac{1}{\lambda_N}v^2\right\}, \tag{6.8}$$

Γ is the gamma function ($\Gamma(1) = 1$):

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \tag{6.9}$$

and Φ is the confluent hypergeometric function:

$$\Phi(\omega, \tau; z) = \frac{\Gamma(\tau)}{\Gamma(\tau - \omega)\Gamma(\omega)} \int_0^1 t^{\omega-1} e^{zt} (1-t)^{\tau-\omega-1} dt, \tag{6.10}$$

and

$$\beta = -\left(j\frac{2v}{\lambda_N}\right)^2 \Big/ \left(4\left(\frac{1}{\lambda_X} + \frac{1}{\lambda_N}\right)\right) = \frac{v^2 \lambda_X}{\lambda_N^2 + \lambda_X \lambda_N}. \tag{6.11}$$

Similarly, the denominator term in (6.2) can be written as:

$$\begin{aligned} &\int_0^\infty \int_0^{2\pi} p(Y|s,\theta)\, p(s,\theta)\, d\theta ds \\ &= C \frac{\Gamma(1)}{2\left(\frac{1}{\lambda_X} + \frac{1}{\lambda_N}\right)} \Phi(1;1;\beta). \end{aligned} \tag{6.12}$$

As in [2, eq. 9.215.1]

$$\Phi(1;1;\beta) = e^\beta. \tag{6.13}$$

Following [2, eq. 9.212.1]

$$\Phi(\alpha+1;1;\beta) = e^\beta \Phi(1-(\alpha+1);1;-\beta) = e^\beta \Phi(-\alpha;1;-\beta). \tag{6.14}$$

The Φ function can be calculated [2, eq. 9.215.1] as:

$$\Phi(\omega, \tau; z) = 1 + \frac{\omega}{\tau}\frac{z}{1!} + \frac{\omega(\omega+1)}{\tau(\tau+1)}\frac{z^2}{2!} + \ldots = \sum_{r=0}^{\infty} \frac{(\omega)_r}{(\tau)_r}\frac{z^r}{r!} \tag{6.15}$$

where

$$(\omega)_r \triangleq \omega \cdot (\omega+1) \ldots (\omega+r-1) \tag{6.16}$$

and $(\omega)_0 \triangleq 1$.

Based on the above calculations of the integrals in (6.2), we can now express the estimated speech as:

$$\hat{s}^{2\alpha} = \frac{\Gamma(\alpha+1)}{\left(\frac{1}{\lambda_X}+\frac{1}{\lambda_N}\right)^{\alpha}}\Phi(-\alpha;1;-\beta) = \frac{\Gamma(\alpha+1)}{\left(\frac{1}{\lambda_X}+\frac{1}{\lambda_N}\right)^{\alpha}}\sum_{r=0}^{\infty}\frac{(-\alpha)_r(-\beta)^r}{(r!)^2} \tag{6.17}$$

In [9], it is suggested that the first 500 terms of the above infinite summation is needed for sufficient accuracy in the approximation.

We simulated the proposed MMSE speech enhancement algorithm with speech signals. The block diagram of such system is given in Figure 6.1. All the simulations are done in the FFT domain with the same noise estimation process. The speech is sampled at 8 kHz and processed on frame-by-frame basis. The frame length used here is 128 samples, which corresponds to 16 ms of speech. The speech can be assumed to be stationary for such frame length. The overlap between successive frames are 64 samples, which represent a 50% overlap. The noise memory is also used here to provide the estimation of noise variance. The input SNR is chosen as 0, 5, 10, 15 and 20 dB. The algorithm is tested with α=0.2, 0.3, 0.5 and 1. α =0.5 is the regular MMSE in spectral domain [41] and α=0.2, 0.3 are the proposed MMSE based on loudness measures. Results are given in Tables 6.1 to 6.6, where the quality of the enhanced speech signal is measured with PESQ, segmental SNR and log-likelihood ratio.

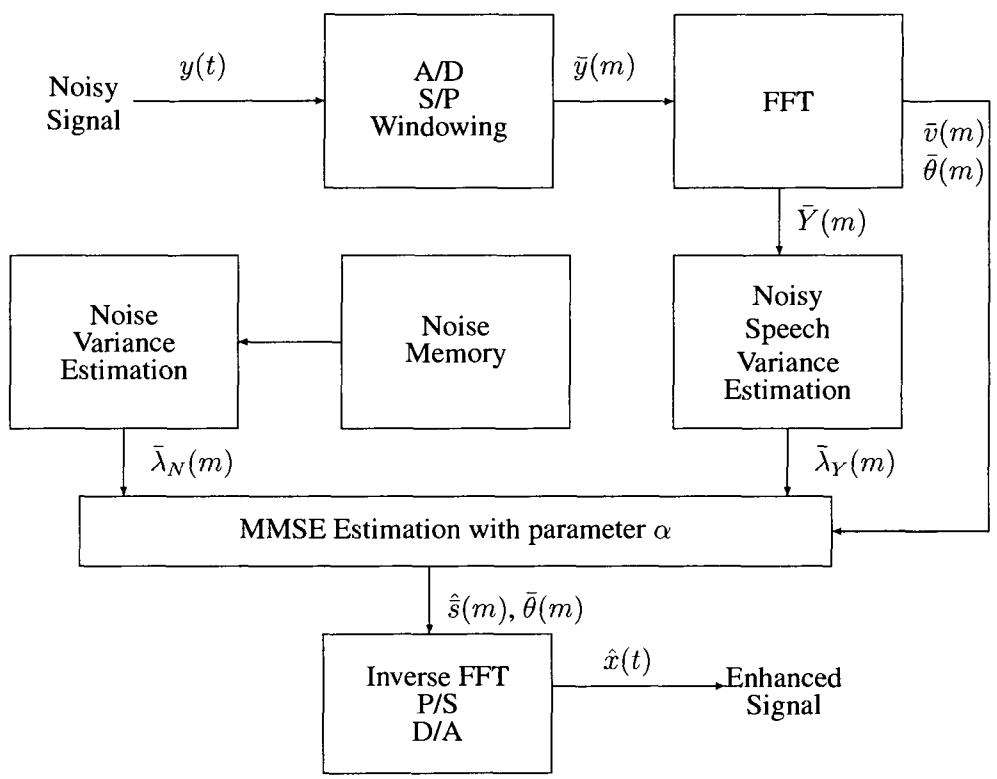

Figure 6.1: The block diagram of the MMSE-based speech enhancement system

Female Speech 1

Input SNR	Input PESQ	PESQ score improvements			
		α=0.2	α=0.3	α=0.5	α=1
0dB	1.36	+0.31	+0.30	+0.27	+0.20
5dB	1.64	+0.34	+0.33	+0.30	+0.21
10dB	1.93	+0.29	+0.27	+0.24	+0.14
15dB	2.32	-0.07	-0.08	-0.11	-0.16
20dB	2.67	-0.61	-0.62	-0.63	-0.66

Female Speech 2

Input SNR	Input PESQ	PESQ score improvements			
		α=0.2	α=0.3	α=0.5	α=1
0dB	1.60	+0.24	+0.23	+0.21	+0.15
5dB	1.87	+0.29	+0.28	+0.26	+0.19
10dB	2.17	+0.15	+0.14	+0.12	+0.06
15dB	2.44	-0.18	-0.19	-0.20	-0.24
20dB	2.73	-0.81	-0.82	-0.83	-0.85

Female Speech 3

Input SNR	Input PESQ	PESQ score improvements			
		α=0.2	α=0.3	α=0.5	α=1
0dB	1.25	+0.43	+0.42	+0.38	+0.27
5dB	1.60	+0.47	+0.45	+0.41	+0.30
10dB	1.96	+0.42	+0.40	+0.36	+0.22
15dB	2.33	+0.04	+0.03	+0.00	-0.09
20dB	2.70	-0.64	-0.65	-0.68	-0.70

Table 6.1: Quality evaluation of various α values for MMSE estimator based on Gaussian speech model estimation: PESQ improvements (Female speech)

Male Speech 1

Input SNR	Input PESQ	PESQ score improvements			
		α=0.2	α=0.3	α=0.5	α=1
0dB	1.39	+0.16	+0.16	+0.15	+0.10
5dB	1.57	+0.33	+0.32	+0.29	+0.21
10dB	1.83	+0.38	+0.37	+0.33	+0.23
15dB	2.15	+0.38	+0.36	+0.33	+0.24
20dB	2.49	+0.06	+0.05	+0.03	-0.02

Male Speech 2

Input SNR	Input PESQ	PESQ score improvements			
		α=0.2	α=0.3	α=0.5	α=1
0dB	1.63	+0.32	+0.31	+0.29	+0.21
5dB	1.95	+0.38	+0.36	+0.34	+0.24
10dB	2.29	+0.38	+0.36	+0.33	+0.23
15dB	2.68	+0.10	+0.09	+0.08	+0.01
20dB	3.11	-0.68	-0.68	-0.69	-0.71

Male Speech 3

Input SNR	Input PESQ	PESQ score improvements			
		α=0.2	α=0.3	α=0.5	α=1
0dB	1.35	+0.35	+0.34	+0.31	+0.22
5dB	1.70	+0.46	+0.44	+0.41	+0.30
10dB	2.06	+0.41	+0.39	+0.35	+0.23
15dB	2.47	+0.12	+0.11	+0.08	-0.02
20dB	2.84	-0.53	-0.54	-0.55	-0.63

Table 6.2: Quality evaluation of various α values for MMSE estimator based on Gaussian speech model estimation: PESQ improvements (Male speech)

Female Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)			
		α=0.2	α=0.3	α=0.5	α=1
0dB	-6.56	2.89	0.04	-0.48	-2.14
5dB	-1.56	3.96	3.78	3.37	2.05
10dB	3.44	6.27	6.14	5.85	4.86
15dB	8.44	7.60	7.53	7.35	6.70
20dB	13.44	7.45	7.41	7.32	6.96

Female Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)			
		α=0.2	α=0.3	α=0.5	α=1
0dB	-11.78	-4.66	-4.94	-5.54	-7.32
5dB	-6.78	-0.71	-0.93	-1.41	-2.89
10dB	-1.78	1.85	1.68	1.29	0.09
15dB	3.22	2.97	2.85	2.58	1.71
20dB	8.22	3.78	3.69	3.49	2.88

Female Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)			
		α=0.2	α=0.3	α=0.5	α=1
0dB	-6.76	-0.16	-0.41	-0.93	-2.54
5dB	-1.76	3.73	3.55	3.14	1.83
10dB	3.24	6.88	6.76	6.47	5.48
15dB	8.24	7.50	7.43	7.25	6.63
20dB	13.24	6.56	6.52	6.41	6.05

Table 6.3: Quality evaluation of various α values for MMSE estimator based on Gaussian speech model estimation: Segmental SNR (Female speech)

Male Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)			
		α=0.2	α=0.3	α=0.5	α=1
0dB	-6.30	5.50	3.08	-0.21	-1.85
5dB	-1.30	4.53	4.34	3.91	2.51
10dB	3.70	8.06	7.92	7.60	6.51
15dB	8.70	8.12	8.03	7.82	7.10
20dB	13.70	5.65	5.60	5.49	5.09

Male Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)			
		α=0.2	α=0.3	α=0.5	α=1
0dB	-5.94	0.29	0.07	-0.39	-1.89
5dB	-0.94	3.99	3.82	3.45	2.26
10dB	4.06	6.89	6.77	6.52	5.67
15dB	9.06	7.63	7.57	7.41	6.87
20dB	14.06	5.19	5.15	5.07	4.79

Male Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)			
		α=0.2	α=0.3	α=0.5	α=1
0dB	-7.87	-1.35	-1.59	-2.10	-3.69
5dB	-2.87	2.37	2.19	1.79	0.50
10dB	2.13	4.81	4.68	4.40	3.45
15dB	7.13	5.86	5.78	5.59	4.95
20dB	12.13	5.71	5.66	5.53	5.16

Table 6.4: Quality evaluation of various α values for MMSE estimator based on Gaussian speech model estimation: Segmental SNR (Male speech)

Female Speech 1

Input SNR	Input LLR	Log Likelihood Ratio			
		α=0.2	α=0.3	α=0.5	α=1
0	2.02	1.56	1.57	1.60	1.70
5	1.66	1.19	1.20	1.22	1.32
10	1.29	0.87	0.88	0.90	1.00
15	0.91	0.63	0.63	0.65	0.74
20	0.60	0.58	0.58	0.60	0.66

Female Speech 2

Input SNR	Input LLR	Log Likelihood Ratio			
		α=0.2	α=0.3	α=0.5	α=1
0	1.61	1.31	1.32	1.33	1.39
5	1.31	1.03	1.03	1.05	1.11
10	1.06	0.84	0.84	0.85	0.89
15	0.78	0.67	0.67	0.68	0.72
20	0.56	0.63	0.63	0.64	0.68

Female Speech 3

Input SNR	Input LLR	Log Likelihood Ratio			
		α=0.2	α=0.3	α=0.5	α=1
0	2.12	1.58	1.60	1.63	1.74
5	1.72	1.16	1.17	1.20	1.32
10	1.31	0.83	0.84	0.86	0.97
15	0.96	0.70	0.70	0.73	0.81
20	0.64	0.59	0.59	0.61	0.68

Table 6.5: Quality evaluation of various α values for MMSE estimator based on Gaussian speech model estimation: Log Likelihood Ratio (Female speech)

Male Speech 1

Input SNR	Input LLR	Log Likelihood Ratio			
		α=0.2	α=0.3	α=0.5	α=1
0	2.49	1.82	1.83	1.87	2.01
5	2.01	1.43	1.45	1.48	1.60
10	1.58	1.09	1.10	1.13	1.23
15	1.17	0.83	0.84	0.86	0.95
20	0.81	0.81	0.83	0.86	0.94

Male Speech 2

Input SNR	Input LLR	Log Likelihood Ratio			
		α=0.2	α=0.3	α=0.5	α=1
0	1.38	1.05	1.06	1.08	1.14
5	1.08	0.76	0.77	0.78	0.85
10	0.78	0.56	0.56	0.57	0.62
15	0.53	0.39	0.39	0.41	0.45
20	0.36	0.38	0.39	0.40	0.43

Male Speech 3

Input SNR	Input LLR	Log Likelihood Ratio			
		α=0.2	α=0.3	α=0.5	α=1
0	1.67	1.38	1.38	1.39	1.45
5	1.35	1.03	1.03	1.05	1.12
10	1.08	0.81	0.81	0.83	0.89
15	0.79	0.60	0.65	0.64	0.67
20	0.55	0.55	0.55	0.57	0.61

Table 6.6: Quality evaluation of various α values for MMSE estimator based on Gaussian speech model estimation: Log Likelihood Ratio (Male speech)

From Tables 6.1 to 6.6, we can see that the MMSE speech enhancement algorithm improves the quality when the input SNR is lower than 10 dB. When the input SNR is high, the MMSE algorithm performs poorly. For all three objective quality measures, the quality is improved for all the α values when the input SNR is below 10 dB.

We will now explain the reason for the degradation in performance for high input SNR cases. Define a gain function G such that the estimated clean signal $\hat{s} = Gv$, where v is the amplitude of the noisy signal FFT. Equation (6.17) can be rewritten as:

$$\hat{s} = \frac{\left(\Gamma\left(\alpha+1\right)\Phi\left(-\alpha;1;-\beta\right)\right)^{\frac{1}{2\alpha}}}{\left(\frac{1}{\lambda_X}+\frac{1}{\lambda_N}\right)^{\frac{1}{2}}} \tag{6.18}$$

Let the *a priori* SNR be given by

$$\xi = \frac{\lambda_X}{\lambda_N} \tag{6.19}$$

and the *a posteriori* SNR be given by

$$\gamma = \frac{v^2}{\lambda_N} \tag{6.20}$$

then we can rewrite β as

$$\beta = \frac{v^2\lambda_X}{\lambda_N^2+\lambda_X\lambda_N} = \frac{\xi}{1+\xi}\gamma. \tag{6.21}$$

Also,

$$\left(\frac{1}{\lambda_X}+\frac{1}{\lambda_N}\right)^{\frac{1}{2}} = \sqrt{\frac{\lambda_X+\lambda_N}{\lambda_X\lambda_N}} = \frac{\frac{v^2}{\lambda_N}}{v\sqrt{\frac{v^2\lambda_X}{\lambda_N^2+\lambda_X\lambda_N}}} = \frac{\gamma}{v\sqrt{\beta}}. \tag{6.22}$$

Substituting for (6.21) and (6.22) in Equation (6.18), we can express the FFT magnitude of the estimated speech as:

$$\hat{s} = \left(\Gamma\left(\alpha+1\right)\Phi\left(-\alpha;1;-\beta\right)\right)^{\frac{1}{2\alpha}}\frac{\sqrt{\beta}}{\gamma}v \tag{6.23}$$

The gain function G is:

$$G_{MMSE,Gaussian} = \left(\Gamma\left(\alpha+1\right)\Phi\left(-\alpha;1;-\beta\right)\right)^{\frac{1}{2\alpha}}\frac{\sqrt{\beta}}{\gamma}. \tag{6.24}$$

for the MMSE estimate and Gaussian assumptions for both speech and noise signals.

Note that G is a gain function of γ and ξ (through the dependence of β on ξ). This relationship is shown in Figure 6.2. ξ is fixed at 15, 5, -5 and -15 dB, while the instantaneous SNR $\gamma - 1$ varies from -15 to 15 dB with 1 dB increment. The gain function for Wiener filter is also depicted in the same figure for comparison.

To understand why the performance of the above algorithm degraded for high input SNRs, we examine the dependence of G on γ and ξ in Figure 6.2. When the prior SNR is high (higher average signal to noise ratio for previous several frames) and γ is low (the amplitude in current frame is low), the gain will be much higher than 0 dB. This means a weak spectral component in a particular frame will be amplified and leads to artifacts in the enhanced signal. This leads to the degradation for the high input SNR cases.

6.2.2 MMSE-Based Speech Enhancement Algorithm with Gaussian Speech Model and Voice Activity Detector

In this section, we address the above concern about the reduced performance for higher SNR. In [41] [114], the probability of speech presence is used to modify the gain function in (6.24). Its main effect is to adjust the gain function when we have weak speech frames within a generally high SNR signal, i.e., the ξ value is high and γ value is low. The MMSE estimation can be modified to:

$$\tilde{s} = \frac{\Lambda(Y)}{1 + \Lambda(Y)} \hat{s}, \tag{6.25}$$

where $\hat{s}$ is the MMSE estimation as given in (6.18) and $\Lambda(y)$ is the likelihood ratio of the speech is present to the speech is absent:

$$\Lambda(y) = \frac{p(y|H_1)}{p(y|H_0)} \tag{6.26}$$

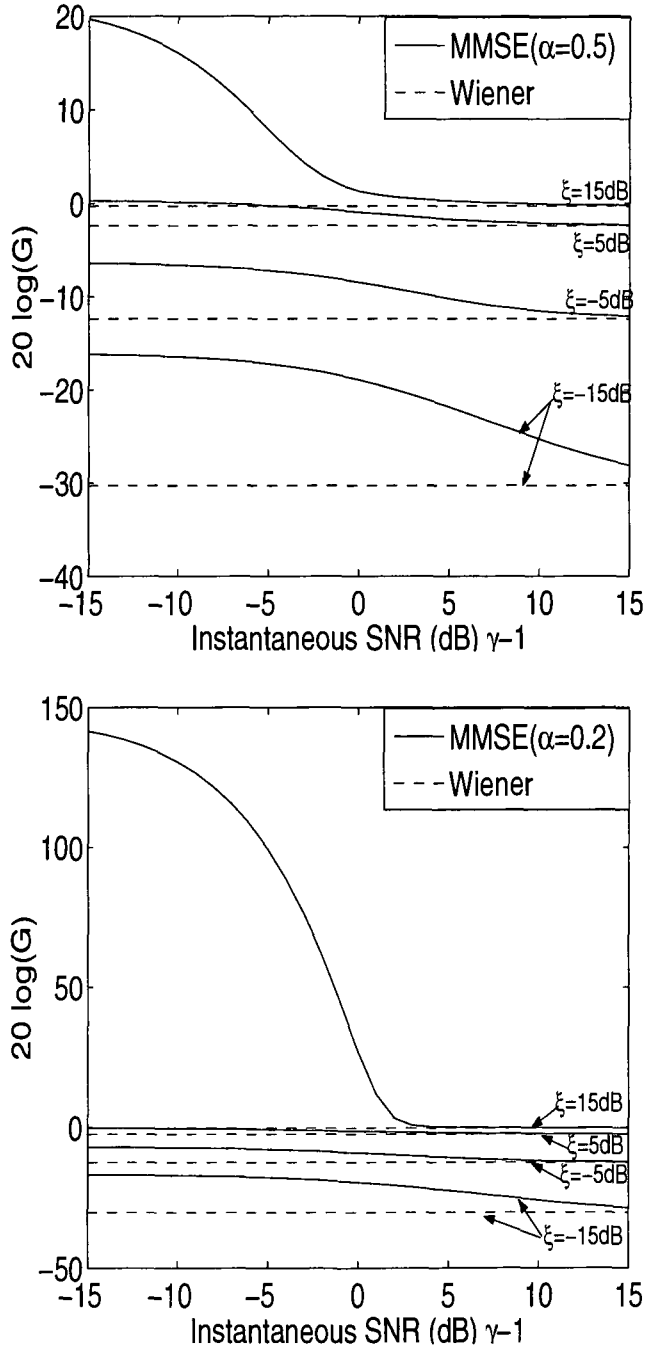

Figure 6.2: The gain function for MMSE as a function of $\gamma - 1$ and ξ for $\alpha = 0.2$ and $\alpha = 0.5$, Gaussian speech model

where H_1 and H_0 represents the two hypothesis that speech is present and absent. We consider the Equation (6.25) as attenuation with a factor of $\Lambda(Y)/(1+\Lambda(Y))$. When the instantaneous SNR is low, the $\Lambda(Y)$ tends to be low and the above factor is much less than 1. Thus the enhanced signal will be attenuated to reduce the artifacts.

As the speech and noise are assumed to be Gaussian with variances λ_X and λ_N, the probabilities in (6.26) can be written as:

$$p(y|H_1) = \frac{1}{\pi(\lambda_X + \lambda_N)} \exp\left\{-\frac{1}{\lambda_X + \lambda_N} y^2\right\} \tag{6.27}$$

and

$$p(y|H_0) = \frac{1}{\pi(\lambda_N)} \exp\left\{-\frac{1}{\lambda_N} y^2\right\}. \tag{6.28}$$

Hence,

$$\Lambda(y) = \frac{p(y|H_1)}{p(y|H_0)} = \frac{\lambda_N}{\lambda_X + \lambda_N} \exp\left\{\frac{\lambda_X}{\lambda_N(\lambda_X + \lambda_N)} y^2\right\} = \frac{1}{1+\xi} \exp\{\beta\}. \tag{6.29}$$

The simulation results of the MMSE-based speech enhancement algorithm with VAD (Gaussian assumption) is summarized in the following Tables 6.7 to 6.12, where again the speech quality is measured using PESQ, segmental SNR and log-likelihood ratio.

Female Speech 1

Input SNR	Input PESQ	PESQ score improvements			
		α=0.2	α=0.3	α=0.5	α=1
0dB	1.36	+0.40	+0.39	+0.39	+0.38
5dB	1.64	+0.50	+0.50	+0.49	+0.47
10dB	1.93	+0.53	+0.53	+0.53	+0.51
15dB	2.32	+0.50	+0.50	+0.50	+0.49
20dB	2.67	+0.50	+0.50	+0.49	+0.48

Female Speech 2

Input SNR	Input PESQ	PESQ score improvements			
		α=0.2	α=0.3	α=0.5	α=1
0dB	1.60	+0.27	+0.27	+0.27	+0.26
5dB	1.87	+0.31	+0.31	+0.31	+0.31
10dB	2.17	+0.32	+0.32	+0.31	+0.31
15dB	2.44	+0.37	+0.37	+0.37	+0.36
20dB	2.73	+0.37	+0.37	+0.36	+0.35

Female Speech 3

Input SNR	Input PESQ	PESQ score improvements			
		α=0.2	α=0.3	α=0.5	α=1
0dB	1.25	+0.62	+0.61	+0.60	+0.58
5dB	1.60	+0.63	+0.63	+0.62	+0.60
10dB	1.96	+0.60	+0.60	+0.59	+0.58
15dB	2.33	+0.52	+0.52	+0.52	+0.51
20dB	2.70	+0.49	+0.49	+0.48	+0.46

Table 6.7: Quality evaluation of various α values for MMSE estimator based on Gaussian speech model estimation (with VAD): PESQ improvements (Female speech)

Male Speech 1

Input SNR	Input PESQ	PESQ score improvements			
		α=0.2	α=0.3	α=0.5	α=1
0dB	1.39	+0.25	+0.25	+0.25	+0.24
5dB	1.57	+0.43	+0.42	+0.42	+0.40
10dB	1.83	+0.49	+0.49	+0.48	+0.47
15dB	2.15	+0.47	+0.47	+0.47	+0.46
20dB	2.49	+0.31	+0.31	+0.31	+0.31

Male Speech 2

Input SNR	Input PESQ	PESQ score improvements			
		α=0.2	α=0.3	α=0.5	α=1
0dB	1.63	+0.35	+0.34	+0.34	+0.33
5dB	1.95	+0.46	+0.46	+0.45	+0.44
10dB	2.29	+0.48	+0.48	+0.46	+0.45
15dB	2.68	+0.51	+0.51	+0.50	+0.50
20dB	3.11	+0.30	+0.31	+0.31	+0.30

Male Speech 3

Input SNR	Input PESQ	PESQ score improvements			
		α=0.2	α=0.3	α=0.5	α=1
0dB	1.35	+0.51	+0.51	+0.50	+0.49
5dB	1.70	+0.56	+0.56	+0.55	+0.53
10dB	2.06	+0.67	+0.66	+0.65	+0.63
15dB	2.47	+0.53	+0.53	+0.53	+0.52
20dB	2.84	+0.48	+0.48	+0.47	+0.46

Table 6.8: Quality evaluation of various α values for MMSE estimator based on Gaussian speech model estimation (with VAD): PESQ improvements (Male speech)

Female Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)			
		α=0.2	α=0.3	α=0.5	α=1
0dB	-6.56	1.99	1.92	1.78	1.45
5dB	-1.56	5.40	5.37	5.29	5.08
10dB	3.44	8.59	8.58	8.55	8.49
15dB	8.44	11.86	11.87	11.89	11.86
20dB	13.44	15.38	15.40	15.44	15.48

Female Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)			
		α=0.2	α=0.3	α=0.5	α=1
0dB	-11.78	-2.36	-2.46	-2.66	-3.10
5dB	-6.78	1.21	1.14	1.00	0.68
10dB	-1.78	4.45	4.41	4.32	4.10
15dB	3.22	7.84	7.82	7.78	7.68
20dB	8.22	11.36	11.36	11.35	11.29

Female Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)			
		α=0.2	α=0.3	α=0.5	α=1
0dB	-6.76	1.51	1.41	1.31	0.98
5dB	-1.76	4.94	4.91	4.83	4.63
10dB	3.24	8.31	8.30	8.26	8.16
15dB	8.24	11.75	11.75	11.75	11.70
20dB	13.24	15.42	15.44	15.46	15.48

Table 6.9: Quality evaluation of various α values for MMSE estimator based on Gaussian speech model estimation (with VAD): Segmental SNR (Female speech)

Male Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)			
		α=0.2	α=0.3	α=0.5	α=1
0dB	-6.30	2.26	2.19	2.07	1.76
5dB	-1.30	5.99	5.94	5.83	5.58
10dB	3.70	9.36	9.34	9.30	9.17
15dB	8.70	12.61	12.62	12.61	12.57
20dB	13.70	15.94	15.96	15.98	16.01

Male Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)			
		α=0.2	α=0.3	α=0.5	α=1
0dB	-5.94	1.68	1.62	1.51	1.23
5dB	-0.94	5.03	5.00	4.94	4.76
10dB	4.06	8.14	8.14	8.12	8.06
15dB	9.06	11.76	11.76	11.77	11.76
20dB	14.06	15.49	15.52	15.56	15.56

Male Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)			
		α=0.2	α=0.3	α=0.5	α=1
0dB	-7.87	-0.78	-0.05	0.64	0.71
5dB	-2.87	3.74	3.71	3.63	3.45
10dB	2.13	6.98	6.97	6.93	6.83
15dB	7.13	10.39	10.39	10.39	10.36
20dB	12.13	14.09	14.11	14.14	14.16

Table 6.10: Quality evaluation of various α values for MMSE estimator based on Gaussian speech model estimation (with VAD): Segmental SNR (Male speech)

Female Speech 1

Input SNR	Input LLR	Log Likelihood Ratio			
		α=0.2	α=0.3	α=0.5	α=1
0	2.02	1.49	1.49	1.49	1.50
5	1.66	1.09	1.10	1.10	1.10
10	1.29	0.79	0.79	0.79	0.79
15	0.91	0.57	0.57	0.57	0.56
20	0.60	0.40	0.40	0.40	0.39

Female Speech 2

Input SNR	Input LLR	Log Likelihood Ratio			
		α=0.2	α=0.3	α=0.5	α=1
0	1.61	1.28	1.28	1.28	1.28
5	1.31	1.03	1.03	1.03	1.03
10	1.06	0.84	0.83	0.83	0.82
15	0.78	0.66	0.66	0.65	0.64
20	0.56	0.46	0.45	0.45	0.44

Female Speech 3

Input SNR	Input LLR	Log Likelihood Ratio			
		α=0.2	α=0.3	α=0.5	α=1
0	2.12	1.40	1.40	1.41	1.45
5	1.72	1.07	1.07	1.07	1.08
10	1.31	0.82	0.82	0.82	0.82
15	0.96	0.62	0.61	0.61	0.61
20	0.64	0.45	0.44	0.44	0.43

Table 6.11: Quality evaluation of various α values for MMSE estimator based on Gaussian speech model estimation (with VAD): Log Likelihood Ratio (Female speech)

Male Speech 1

Input SNR	Input LLR	Log Likelihood Ratio			
		α=0.2	α=0.3	α=0.5	α=1
0	2.49	1.71	1.71	1.72	1.73
5	2.01	1.29	1.30	1.30	1.32
10	1.58	1.07	1.07	1.07	1.08
15	1.17	0.80	0.80	0.86	0.79
20	0.81	0.60	0.60	0.60	0.59

Male Speech 2

Input SNR	Input LLR	Log Likelihood Ratio			
		α=0.2	α=0.3	α=0.5	α=1
0	1.38	0.98	0.98	0.99	0.99
5	1.08	0.78	0.78	0.77	0.77
10	0.78	0.61	0.61	0.61	0.60
15	0.53	0.42	0.42	0.41	0.41
20	0.36	0.32	0.31	0.31	0.31

Male Speech 3

Input SNR	Input LLR	Log Likelihood Ratio			
		α=0.2	α=0.3	α=0.5	α=1
0	1.67	1.33	1.33	1.33	1.33
5	1.35	1.06	1.06	1.06	1.05
10	1.08	0.84	0.83	0.83	0.83
15	0.79	0.59	0.59	0.59	0.57
20	0.55	0.42	0.41	0.41	0.40

Table 6.12: Quality evaluation of various α values for MMSE estimator based on Gaussian speech model estimation (with VAD): Log Likelihood Ratio (Male speech)

From Tables 6.7 to 6.12, we can see that the algorithm now provides consistent improvement over the full range of SNR of the original noisy speech. The voice activity detector is thus shown to have helped to further improve the speech quality. It reduces the enormously large amplification of the weak speech frames in high prior SNR cases. It is also clear that the performance is insensitive to the choice of α except for segmental SNR which shows a slight improvement for smaller αs

6.2.3 MMSE-Based Speech Enhancement Algorithm for Laplacian Speech Model

In [30], the Laplacian model is used for speech signals. The real and imaginary part of the speech spectra are assumed to be jointly Laplacian distributed. It is shown that the Laplacian-based algorithm can provide consistent improvement over the Gaussian-based algorithm, especially when the input SNR is low. The overall PESQ improvement is between 0 and 0.24. The overall Segmental SNR improvement is 0.6-1.7 dB.

The MMSE estimator for both the Gaussian and Laplacian speech model used the Hypergeometric function. The hypergeometric functions can be expressed as a sum of an infinite number of terms. Even though a reasonable approximation can be achieved with 500 terms for the hypergeometric functions, the calculational load is huge. For the Laplacian speech model, a series of 40 hypergeometric functions is needed at each sample, the computational load is even higher. This leads us to explore for a substitute of the MMSE approach with the aim of achieving comparable performance with less computations.

6.3 Proposed ML-Based Speech Enhancement Algorithm with Gaussian Speech Model

In this section, we use the Maximum Likelihood (ML) estimation as the substitute of the MMSE estimation aiming at reduced complexity. Both the Gaussian and Laplacian statistical models are used for speech signal modelling and the results are compared.

6.3.1 The Likelihood Function under Gaussian Speech Model and its Maximization

In this section, we assume that the speech component $X = se^{j\theta}$ has a zero-mean Gaussian distribution with variance λ_X and the noise N has a zero-mean Gaussian distribution with variance λ_N. Also, the noisy speech component is $Y = ve^{j\kappa}$. Following these assumptions, the distribution functions can be written as:

$$f_X(X) = \frac{1}{\pi\lambda_X} e^{-\frac{s^2}{\lambda_X}} \tag{6.30}$$

and

$$f_N(Y|X) = \frac{1}{\pi\lambda_N} e^{-\frac{|Y-X|^2}{\lambda_N}}. \tag{6.31}$$

Let $t = s^{2\alpha}$, which is proportional to the loudness measure of X. The PDF of t can be derived as following:

$$f_T(t) = \frac{f_X(t^{\frac{1}{2\alpha}})}{\left|\frac{d(X^{2\alpha})}{dX}\right|_{|x|=t^{\frac{1}{2\alpha}}}} = \frac{f_X(t^{\frac{1}{2\alpha}})}{2\alpha t^{\frac{2\alpha-1}{2\alpha}}} = \frac{e^{-\frac{t^{\frac{1}{\alpha}}}{\lambda_X}}}{2\alpha t^{\frac{2\alpha-1}{2\alpha}} \cdot \pi\lambda_X}, \tag{6.32}$$

assuming $t \neq 0$.

In ML estimation, we want to maximize the probability of the loudness measure given the observation, i.e., $f(t|Y)$. From the property of conditional probability

$$f(t|Y) = f(t,Y)/f(Y) \tag{6.33}$$

If we assume Y is known, maximizing $f(t|Y)$ is equivalent to maximizing $f(t,Y)$. Since t is a function of X only, we assume that t is independent with N for simplicity, then:

$$f_{t,Y}(t,Y) = f_t(t) f_N(Y-X) = \frac{e^{-\frac{t^{\frac{1}{\alpha}}}{\lambda_X}}}{2\alpha t^{\frac{2\alpha-1}{2\alpha}} \cdot \pi\lambda_X} \cdot \frac{1}{\pi\lambda_N} e^{-\frac{|Y-X|^2}{\lambda_N}} \tag{6.34}$$

For fixed Y and t values, the above $f(t,Y)$ is maximized when $f_N(Y-X)$ is maximized, i.e., $|Y-X|$ is minimized. When t is fixed, the amplitude of X is fixed and $|Y-X|$ is minimized by choosing X have the same phase as Y. This means the estimate of X should always have the same phase with Y. In the following, only the enhancement of the amplitude is considered.

Let C be the constant in (6.34)

$$C = \frac{1}{2\alpha \cdot \pi^2 \lambda_X \lambda_N}. \tag{6.35}$$

Differentiate (6.34) with respect to t and equating to 0, we get:

$$C \cdot \frac{t^{\frac{2\alpha-1}{2\alpha}} \left(-\frac{\frac{1}{\alpha} t^{\left(\frac{1}{\alpha}-1\right)}}{\lambda_X} - \frac{\frac{1}{\alpha} t^{\left(\frac{1}{\alpha}-1\right)} - 2v\frac{1}{2\alpha} t^{\left(\frac{1}{2\alpha}-1\right)}}{\lambda_N} \right) - \left(\frac{2\alpha-1}{2\alpha}\right) t^{-\frac{1}{2\alpha}}}{\left(t^{\frac{2\alpha-1}{2\alpha}}\right)^2} \cdot e^{-\frac{t^{\frac{1}{\alpha}}}{\lambda_X} - \frac{(v-t^{\frac{1}{\alpha}})^2}{\lambda_N}} = 0 \tag{6.36}$$

As $e^{-\frac{t^{\frac{1}{\alpha}}}{\lambda_X} - \frac{(v-t^{\frac{1}{\alpha}})^2}{\lambda_N}} \neq 0$ and $t \neq 0$, then

$$t^{\frac{2\alpha-1}{2\alpha}} \left(-\frac{\frac{1}{\alpha} t^{\left(\frac{1}{\alpha}-1\right)}}{\lambda_X} - \frac{\frac{1}{\alpha} t^{\left(\frac{1}{\alpha}-1\right)} - 2v\frac{1}{2\alpha} t^{\left(\frac{1}{2\alpha}-1\right)}}{\lambda_N} \right) - \left(\frac{2\alpha-1}{2\alpha}\right) t^{-\frac{1}{2\alpha}} = 0. \tag{6.37}$$

Simplifying the above equation by multiplying it by $2\lambda_X\lambda_N\alpha$, we get:

$$2t^{\frac{-1}{2\alpha}} \left(-t^{\frac{1}{\alpha}}\lambda_N - t^{\frac{1}{\alpha}}\lambda_X + v\lambda_X t^{\frac{1}{2\alpha}} \right) - (2\alpha-1)\lambda_X\lambda_N t^{-\frac{1}{2\alpha}} = 0. \tag{6.38}$$

$$2\left(\lambda_X+\lambda_N\right)t^{\frac{1}{\alpha}}-2v\lambda_X t^{\frac{1}{2\alpha}}+\left(2\alpha-1\right)\lambda_X\lambda_N=0, \tag{6.39}$$

Recall that $s=t^{\frac{1}{2\alpha}}$, then the optimum ML estimate $\hat{s}$ is given by the solution to:

$$2\left(\lambda_X+\lambda_N\right)\hat{s}^2-2v\lambda_X\hat{s}+\left(2\alpha-1\right)\lambda_X\lambda_N=0. \tag{6.40}$$

This is a quadratic equation with the discriminant:

$$\Delta=v^2\lambda_X^2+2(1-2\alpha)\lambda_X\lambda_N\left(\lambda_X+\lambda_N\right). \tag{6.41}$$

When $\Delta<0$, (6.38) is negative for all t values. So the first derivative of (6.34) is always negative, which means (6.34) is strictly decreasing. In other words, (6.34) is maximized when $t=0$.

When $\Delta\geq 0$, the solution is

$$\hat{s}_{1,2}=\frac{v\lambda_X\pm\sqrt{v^2\lambda_X^2+2(1-2\alpha)\lambda_X\lambda_N\left(\lambda_X+\lambda_N\right)}}{2\left(\lambda_X+\lambda_N\right)}. \tag{6.42}$$

$\hat{s}_{1,2}$ correspond to either maximum or minimum of the (6.34). The second derivative of (6.34) is needed to determine which is which.

The first derivative in (6.36) is rewritten as:

$$\frac{d}{dt}f_{t,Y}\left(t,Y\right)=C'F_1(t)F_2(t) \tag{6.43}$$

where

$$C'=C\frac{1}{2\lambda_X\lambda_N\alpha} \tag{6.44}$$

$$F_1(t)=\frac{t^{\frac{-1}{2\alpha}}}{\left(t^{\frac{2\alpha-1}{2\alpha}}\right)^2}\cdot e^{-\frac{t^{\frac{1}{\alpha}}}{\lambda_X}-\frac{(v-t^{\frac{1}{\alpha}})^2}{\lambda_N}} \tag{6.45}$$

and

$$F_2(t)=2\left(-t^{\frac{1}{\alpha}}\lambda_N-t^{\frac{1}{\alpha}}\lambda_X+v\lambda_X t^{\frac{1}{2\alpha}}\right)-\left(2\alpha-1\right)\lambda_X\lambda_N \tag{6.46}$$

The second derivative of $f_{t,Y}(t,Y)$ at $t=\hat{s}_{1,2}^{2\alpha}$ is given by:

$$\begin{aligned}\frac{d^2}{dt^2}f_{t,Y}(t,Y)\Big|_{t=\hat{s}_{1,2}^{2\alpha}} &= \frac{d}{dt}\left[C'F_1(t)F_2(t)\right]\Big|_{t=\hat{s}_{1,2}^{2\alpha}} \\ &= \left[C'F_1(t)\frac{d}{dt}F_2(t)+C'F_1(t)\frac{d}{dt}F_2(t)\right]\Big|_{t=\hat{s}_{1,2}^{2\alpha}}. \end{aligned} \tag{6.47}$$

Note that for $\hat{s}=\hat{s}_{1,2}$, $F_1(t)\cdot F_2(t)$=0. Since we can assume that t, λ_X, λ_N and α to be all positive, $F_1(t)$ and C' are always positive. Thus,

$$F_2(t)|_{t=\hat{s}_{1,2}^{2\alpha}}=0, \tag{6.48}$$

$$\frac{d^2}{dt^2}f_{t,Y}(t,Y)\Big|_{t=\hat{s}_{1,2}^{2\alpha}} = \left[C'F_1(t)\frac{d}{dt}F_2(t)\right]\Big|_{t=\hat{s}_{1,2}^{2\alpha}}. \tag{6.49}$$

As we have $C'\geq 0$, $F_1(t)\geq 0$.

$$\begin{aligned}\frac{d}{dt}F_2(t)\Big|_{t=\hat{s}_{1,2}^{2\alpha}} &= \left[\frac{1}{\alpha}\left(-2\lambda_X-2\lambda_N\right)t^{\frac{1}{\alpha}-1}+\frac{1}{2\alpha}2v\lambda_X t^{\frac{1}{2\alpha}-1}\right]\Big|_{t=\hat{s}_{1,2}^{2\alpha}} \\ &= \left[\frac{\left(-4\lambda_X-4\lambda_N\right)t^{\frac{1}{\alpha}}+2v\lambda_X t^{\frac{1}{2\alpha}}}{2\alpha t}\right]\Big|_{t=\hat{s}_{1,2}^{2\alpha}} \\ &= \left[\frac{\left(-4\lambda_X-4\lambda_N\right)s^2+2v\lambda_X s}{2\alpha s^{2\alpha}}\right]\Big|_{s=\hat{s}_{1,2}} \end{aligned} \tag{6.50}$$

Assign $\hat{s}_1$ to the root with the + sign.

$$\begin{aligned}\frac{d}{dt}F_2(t)\Big|_{t=\hat{s}_1^{2\alpha}} &= \frac{\left(-4\lambda_X-4\lambda_N\right)\left(\frac{v\lambda_X+\sqrt{\Delta}}{2(\lambda_X+\lambda_N)}\right)^2+2v\lambda_X\left(\frac{v\lambda_X+\sqrt{\Delta}}{2(\lambda_X+\lambda_N)}\right)}{2\alpha s^{2\alpha}} \\ &= \frac{-2\left(v\lambda_X+\sqrt{\Delta}\right)^2+2v\lambda_X\left(v\lambda_X+\sqrt{\Delta}\right)}{4\alpha s^{2\alpha}\left(\lambda_X+\lambda_N\right)} \\ &= \frac{-v\lambda_X\sqrt{\Delta}-\Delta}{4\alpha s^{2\alpha}\left(\lambda_X+\lambda_N\right)}<0 \end{aligned} \tag{6.51}$$

which means $\hat{s}_1$ corresponding to the maximum of $f_{t,Y}(t,Y)$.

Also,

$$\left.\frac{d}{dt}F_2(t)\right|_{t=\hat{s}_2^{2\alpha}} = \frac{-2\left(v\lambda_X - \sqrt{\Delta}\right)^2 + 2v\lambda_X\left(v\lambda_X - \sqrt{\Delta}\right)}{4\alpha s^{2\alpha}\left(\lambda_X + \lambda_N\right)}$$
$$= \frac{\sqrt{\Delta}v\lambda_X - \sqrt{\Delta}}{4\alpha s^{2\alpha}\left(\lambda_X + \lambda_N\right)} \tag{6.52}$$

There are two possibilities, if $v\lambda_X - \sqrt{\Delta} \geq 0$, then the above

$$\left.\frac{d}{dt}F_2(t)\right|_{t=\hat{s}_2^{2\alpha}} \geq 0 \tag{6.53}$$

which means $\hat{s}_2$ corresponding to the minimum of $f_{t,Y}(t, Y)$. If $v\lambda_X - \sqrt{\Delta} < 0$, the root $\hat{s}_2 < 0$, which will be discarded since s is an estimation of amplitude and can't be negative. Either way, $\hat{s}_1$ corresponds to the only maximum of $f_{t,Y}(t, Y)$.

Overall, the maximum likelihood under Gaussian assumption is given by $\hat{s}$, which can be expressed as:

$$\hat{s}_{ML,G} = \frac{v\lambda_X + \sqrt{\Delta}}{2\left(\lambda_X + \lambda_N\right)}. \tag{6.54}$$

when

$$\Delta = v^2\lambda_X^2 + 2(1-2\alpha)\lambda_X\lambda_N\left(\lambda_X + \lambda_N\right) \geq 0 \tag{6.55}$$

When $\Delta < 0$

$$\hat{s}_{ML,G} = 0. \tag{6.56}$$

A block diagram of the ML-based speech enhancement algorithms is shown in Figure 6.3.

The noise variance can be estimated as following:

$$\widehat{\lambda_N}(m) = \beta_N\widehat{\lambda_N}(m-1) + (1-\beta_N)|N(m)|^2, \tag{6.57}$$

and the β_N is chosen to set the time constant of above adaptive process to be 0.5 second, in which the variation of noise variance is negligible. $N(m)$ is the FFT components of the noise reference signal at time m that was taken from the noise memory.

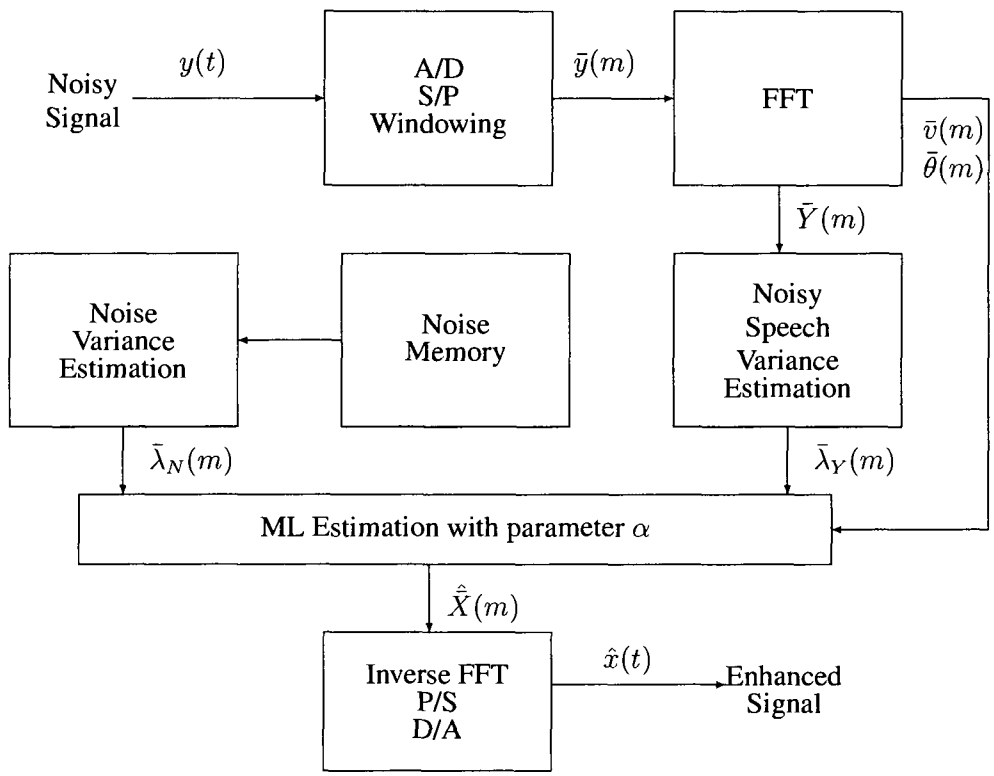

Figure 6.3: The block diagram of the ML-based speech enhancement system

The speech variance should be estimated as $\max(\lambda_Y - \lambda_N, 0)$ under noisy environments. The variance of the noisy speech signal λ_Y should be estimated as:

$$\widehat{\lambda_Y}(m) = \beta_Y \widehat{\lambda_Y}(m-1) + (1-\beta_Y)|Y(m)|^2, \tag{6.58}$$

and the β_Y is chosen to set the time constant of above adaptive process to be 10 ms, in which the speech can be assumed to be stationary.

6.3.2 Simulation Results for Proposed ML-based Algorithm with Gaussian Speech Model

In this section, the proposed speech enhancement algorithm based on the ML estimation is simulated. The speech signal amplitude components in the FFT domain will be enhanced with the proposed ML estimation. The phase components of the noisy signal are retained and used for the enhanced signal. The enhanced speech signal is then recovered by applying IFFT to these enhanced components. The complete speech enhancement algorithm is shown in Table 6.3.2

But first, the results of the proposed speech enhancement algorithm are shown with ideal variance estimation. The speech signal variance λ_X is estimated with the clean speech:

$$\widehat{\lambda_X}(m) = \beta_X \widehat{\lambda_X}(m-1) + (1-\beta_X)|X(m)|^2, \tag{6.59}$$

and the β_X is chosen to set the time constant of the above adaptive process to be 10 ms. $X(m)$ is the FFT components of the clean speech signal at frame m. Under such assumption, the optimal performances that can be achieved with this algorithm are shown.

We tested the algorithms using different α values: 0.2, 0.3, 0.5, 1 and 1.5. The quality of the enhanced signals (as measured by PESQ segmental SNR and log-likelihood ratio) is shown in the following Tables 6.14 to 6.19.

β_Y is chosen to let the timeconstant of the adaptive process to be 10 msec.

β_N is chosen to let the time constant of the adaptive process to be 0.5 sec.

For each time step m do,

$$\begin{aligned}
\bar{Y}(m) &= FFT\left(\bar{y}(m)\right) \\
\bar{n}(m) &\leftarrow \text{Read from noise memory} \\
\bar{N}(m) &= FFT\left(\bar{n}(m)\right) \\
\hat{V}(m) &= [v_1(m)|v_2(m)|\ldots|v_R(m)] = \left|\bar{Y}(m)\right|; \\
\hat{\theta}(m) &= [\theta_1(m)|\theta_2(m)|\ldots|\theta_R(m)] = \angle\bar{Y}(m); \\
\hat{U}(m) &= [u_1(m)|u_2(m)|\ldots|u_R(m)] = \left|\bar{N}(m)\right|;
\end{aligned}$$

For i= 1, 2,. . ., R do

$$\begin{aligned}
\lambda_{Yi}(m) &= \beta_Y\lambda_{Yi}(m-1) + (1-\beta_Y)v_i^2(m) \\
\lambda_{Ni}(m) &= \beta_N\lambda_{Ni}^2(m-1) + (1-\beta_N)u_i^2(m) \\
\lambda_{Xi}(m) &= \max\left(\lambda_{Yi}(m) - \lambda_{Ni}(m), 0\right) \\
\Delta_i(m) &= v_i(m)^2\lambda_{Xi}(m)^2 + 2(1-2\alpha)\lambda_{Xi}(m)\lambda_{Ni}(m)\lambda_{Yi}(m)
\end{aligned}$$

When $\Delta_i(m) \geq 0$

$$\hat{s}_i(m) = \frac{v_i(m)\lambda_{Xi}(m) + \sqrt{\Delta_i(m)}}{2\lambda_{Yi}(m)}.$$

When $\Delta_i(m) < 0$

$$\hat{s}_i(m) = 0$$

end;

$$\begin{aligned}
\hat{S}(m) &= [\hat{s}_1(m)|\hat{s}_2(m)|\ldots|\hat{s}_R(m)] \\
\hat{X}(m) &= \hat{S}(m)\cdot e^{j\hat{\theta}(m)} \\
\hat{x}(m) &= IFFT\left(\hat{X}(m)\right)
\end{aligned}$$

Table 6.13: Proposed ML-based Speech Enhancement Algorithm (Gaussian Speech Model)

Female Speech 1

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.36	+1.06	+1.08	+1.05	+0.97	+0.97
5dB	1.64	+0.96	+0.98	+1.00	+0.88	+0.88
10dB	1.93	+0.82	+0.85	+0.89	+0.81	+0.81
15dB	2.32	+0.71	+0.74	+0.78	+0.71	+0.69
20dB	2.67	+0.58	+0.62	+0.68	+0.66	+0.64

Female Speech 2

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.60	+0.90	+0.91	+0.86	+0.72	+0.71
5dB	1.87	+0.83	+0.84	+0.83	+0.73	+0.73
10dB	2.17	+0.70	+0.72	+0.70	+0.59	+0.59
15dB	2.44	+0.59	+0.62	+0.62	+0.59	+0.58
20dB	2.73	+0.48	+0.50	+0.53	+0.51	+0.51

Female Speech 3

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.25	+1.32	+1.34	+1.32	+1.18	+1.18
5dB	1.60	+1.16	+1.19	+1.21	+1.08	+1.08
10dB	1.96	+0.94	+0.97	+1.04	+0.96	+0.95
15dB	2.33	+0.74	+0.76	+0.81	+0.78	+0.77
20dB	2.70	+0.49	+0.53	+0.57	+0.63	+0.63

Table 6.14: Quality evaluation of various α values for ML estimator based on Gaussian speech model with ideal variance estimation: PESQ improvements (Female speech)

Male Speech 1

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.39	+0.99	+1.00	+0.84	+0.65	+0.63
5dB	1.57	+0.89	+0.91	+0.86	+0.66	+0.63
10dB	1.83	+0.79	+0.81	+0.80	+0.63	+0.60
15dB	2.15	+0.61	+0.64	+0.67	+0.57	+0.55
20dB	2.49	+0.44	+0.47	+0.51	+0.47	+0.45

Male Speech 2

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.63	+0.95	+0.97	+0.94	+0.87	+0.87
5dB	1.95	+0.91	+0.93	+0.92	+0.91	+0.82
10dB	2.29	+0.71	+0.73	+0.71	+0.66	+0.63
15dB	2.68	+0.67	+0.70	+0.68	+0.60	+0.58
20dB	3.11	+0.40	+0.43	+0.48	+0.37	+0.34

Male Speech 3

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.35	+1.31	+1.33	+1.31	+1.15	+1.15
5dB	1.70	+1.14	+1.17	+1.20	+1.07	+1.07
10dB	2.06	+0.93	+0.96	+1.00	+0.94	+0.92
15dB	2.47	+0.71	+0.74	+0.83	+0.82	+0.80
20dB	2.84	+0.52	+0.54	+0.61	+0.63	+0.61

Table 6.15: Quality evaluation of various α values for ML estimator based on Gaussian speech model with ideal variance estimation: PESQ improvements (Male speech)

Female Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.56	4.77	4.97	5.29	4.93	4.67
5dB	-1.56	7.13	7.31	7.59	7.23	6.95
10dB	3.44	9.75	9.90	10.13	9.73	9.42
15dB	8.44	12.86	13.01	13.22	12.84	12.50
20dB	13.44	16.24	16.36	16.53	16.19	15.79

Female Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-11.78	2.72	3.01	3.74	3.59	3.42
5dB	-6.78	4.57	4.84	5.44	5.32	5.11
10dB	-1.78	6.85	7.09	7.60	7.48	7.25
15dB	3.22	9.30	9.50	9.91	9.74	9.50
20dB	8.22	12.39	12.55	12.88	12.74	12.48

Female Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.76	4.41	4.64	5.10	4.89	4.60
5dB	-1.76	6.71	6.92	7.31	7.14	6.87
10dB	3.24	9.41	9.59	9.93	9.78	9.49
15dB	8.24	12.56	12.72	12.99	12.84	12.57
20dB	13.24	16.05	16.18	16.40	16.24	15.93

Table 6.16: Quality evaluation of various α values for ML estimator based on Gaussian speech model with ideal variance estimation: Segmental SNR (Female speech)

Male Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.30	5.58	5.78	6.11	5.73	5.43
5dB	-1.30	8.10	8.28	8.56	8.25	7.99
10dB	3.70	10.82	10.99	11.24	10.87	10.58
15dB	8.70	13.69	13.84	14.08	13.74	13.46
20dB	13.70	17.01	17.14	17.33	16.98	16.65

Male Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-5.94	4.43	4.65	5.10	4.89	4.67
5dB	-0.94	6.83	7.04	7.44	7.27	6.98
10dB	4.06	9.49	9.66	10.01	9.75	9.42
15dB	9.06	12.58	12.71	12.95	12.74	12.45
20dB	14.06	16.21	16.32	16.52	16.34	15.05

Male Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-7.87	3.64	3.90	4.50	4.32	4.13
5dB	-2.87	5.78	6.01	6.50	6.30	6.08
10dB	2.13	8.42	8.61	8.96	8.75	8.49
15dB	7.13	11.34	11.49	11.76	11.50	11.23
20dB	12.13	14.86	14.99	15.22	15.04	14.68

Table 6.17: Quality evaluation of various α values for ML estimator based on Gaussian speech model with ideal variance estimation: Segmental SNR (Male speech)

Female Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.02	0.44	0.44	0.65	0.74	0.74
5	1.66	0.41	0.41	0.56	0.66	0.66
10	1.29	0.36	0.36	0.45	0.55	0.55
15	0.91	0.30	0.30	0.33	0.42	0.43
20	0.60	0.25	0.24	0.26	0.32	0.33

Female Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.61	0.44	0.44	0.63	0.67	0.67
5	1.31	0.42	0.42	0.56	0.61	0.62
10	1.06	0.38	0.38	0.52	0.57	0.57
15	0.78	0.31	0.32	0.43	0.49	0.49
20	0.56	0.25	0.25	0.32	0.39	0.40

Female Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.12	0.47	0.47	0.66	0.74	0.76
5	1.72	0.46	0.45	0.58	0.67	0.68
10	1.31	0.42	0.41	0.49	0.60	0.60
15	0.96	0.39	0.38	0.42	0.50	0.52
20	0.64	0.31	0.31	0.34	0.40	0.42

Table 6.18: Quality evaluation of various α values for ML estimator based on Gaussian speech model with ideal variance estimation: Log Likelihood Ratio (Female speech)

Male Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.49	0.61	0.60	0.72	0.78	0.79
5	2.01	0.56	0.55	0.64	0.71	0.71
10	1.58	0.53	0.52	0.57	0.63	0.63
15	1.17	0.46	0.45	0.49	0.56	0.57
20	0.81	0.38	0.38	0.38	0.44	0.45

Male Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.38	0.40	0.41	0.59	0.67	0.67
5	1.08	0.37	0.38	0.51	0.62	0.60
10	0.78	0.34	0.35	0.44	0.53	0.53
15	0.53	0.30	0.30	0.36	0.43	0.44
20	0.36	0.24	0.24	0.28	0.34	0.36

Male Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.67	0.48	0.48	0.67	0.75	0.75
5	1.35	0.44	0.45	0.59	0.69	0.69
10	1.08	0.39	0.38	0.47	0.57	0.57
15	0.79	0.35	0.35	0.40	0.47	0.47
20	0.55	0.29	0.29	0.31	0.37	0.38

Table 6.19: Quality evaluation of various α values for ML estimator based on Gaussian speech model with ideal variance estimation: Log Likelihood Ratio (Male speech)

The results are summarized in tables 6.14 to 6.19, we clearly see that:

• All the tested α values provide higher PESQ, higher segmental SNR and lower IIR, i.e., the speeches are enhanced for all the tests.

• From the five selected α values, $\alpha = 0.3$ provides the best PESQ improvements when the input SNR is low. When the input SNR is high, α=0.5 will outperform $\alpha = 0.3$ slightly. Overall, the three α values of 0.2, 0.3 and 0.5 lead to similar PESQ improvements. The differences between these three are less than 0.1 generally. The PESQ decreases noticeably when α is 1 or 1.5.

• For the segmental SNRs for the enhanced signals, α=0.5 provides the best segmental SNR improvements. α=0.3 has slightly worse segmental SNR scores. For all five tested α values, the segmental SNR is within range of 1 dB comparing to the overall improvements.

• For the distortion measure LLR, α=0.2 and 0.3 provide comparable results. And the LLR increases as α increases beyond 0.3. This measure varies more significantly than the previous two. α=1.5 can lead to 20-60 % more distortion overall measured by LLR.

• For the subjective listening tests, the larger α values leads to better noise removal as well as higher distortion. When the input SNR is high, results from α=0.3 or 0.5 are comparable. When the input SNR is low, α=0.3 performs better. α=0.3 provide lower distortion to the speech itself, while α=0.5 removes more noise. When the input SNR is low, α=0.5 leads to clear damage to the speech itself. So α=0.3 is more favorable overall, which is agreeable with the loudness model.

Next, the simulations are done with actual speech variance estimation from the noisy signal. The noise variance is subtracted from the noisy speech variance as the estimate of the speech variance.

Female Speech 1

Input SNR	Input LLR	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.36	+0.23	+0.24	+0.26	+0.30	+0.08
5dB	1.64	+0.23	+0.25	+0.28	+0.36	+0.33
10dB	1.93	+0.28	+0.29	+0.33	+0.41	+0.19
15dB	2.32	+0.27	+0.29	+0.33	+0.42	+0.34
20dB	2.67	+0.24	+0.26	+0.31	+0.41	+0.38

Female Speech 2

Input SNR	Input LLR	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.60	+0.17	+0.18	+0.19	+0.16	-0.25
5dB	1.87	+0.21	+0.22	+0.24	+0.22	-0.19
10dB	2.17	+0.19	+0.20	+0.22	+0.20	-0.09
15dB	2.44	+0.22	+0.23	+0.25	+0.26	+0.19
20dB	2.73	+0.20	+0.21	+0.24	+0.26	+0.21

Female Speech 3

Input SNR	Input LLR	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.25	+0.28	+0.30	+0.34	+0.42	+0.29
5dB	1.60	+0.29	+0.31	+0.36	+0.48	+0.38
10dB	1.96	+0.30	+0.32	+0.36	+0.47	+0.45
15dB	2.33	+0.32	+0.34	+0.38	+0.49	+0.45
20dB	2.70	+0.25	+0.27	+0.31	+0.45	+0.34

Table 6.20: Quality evaluation of various α values for ML estimator based on Gaussian speech model: PESQ improvements (Female speech)

Male Speech 1

Input SNR	Input LLR	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.39	+0.09	+0.09	+0.10	+0.11	-0.05
5dB	1.57	+0.20	+0.21	+0.23	+0.32	+0.16
10dB	1.83	+0.28	+0.29	+0.33	+0.39	+0.23
15dB	2.15	+0.25	+0.26	+0.30	+0.35	+0.26
20dB	2.49	+0.21	+0.23	+0.26	+0.27	+0.19

Male Speech 2

Input SNR	Input LLR	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.63	+0.18	+0.19	+0.21	+0.26	+0.05
5dB	1.95	+0.24	+0.26	+0.29	+0.35	+0.21
10dB	2.29	+0.25	+0.27	+0.31	+0.35	+0.28
15dB	2.68	+0.32	+0.34	+0.38	+0.40	+0.27
20dB	3.11	+0.17	+0.18	+0.22	+0.26	+0.21

Male Speech 3

Input SNR	Input LLR	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.35	+0.28	+0.29	+0.33	+0.39	+0.29
5dB	1.70	+0.34	+0.35	+0.39	+0.51	+0.32
10dB	2.06	+0.35	+0.37	+0.43	+0.53	+0.53
15dB	2.47	+0.35	+0.37	+0.42	+0.48	+0.34
20dB	2.84	+0.28	+0.29	+0.34	+0.42	+0.34

Table 6.21: Quality evaluation of various α values for ML estimator based on Gaussian speech model: PESQ improvements (Male speech)

Female Speech 1

Input SNR	Input LLR	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.56	-1.89	-1.60	-0.90	1.11	2.11
5dB	-1.56	2.39	2.64	3.19	4.42	4.77
10dB	3.44	6.53	6.73	7.16	7.76	7.69
15dB	8.44	10.72	10.88	11.21	11.31	10.87
20dB	13.44	14.89	15.01	15.22	14.94	14.26

Female Speech 2

Input SNR	Input LLR	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-11.78	-7.13	-6.83	-6.12	-3.80	-2.01
5dB	-6.78	-2.59	-2.32	-1.68	0.18	1.52
10dB	-1.78	1.62	1.84	2.34	3.49	4.39
15dB	3.22	5.83	6.02	6.41	7.06	7.44
20dB	8.22	10.06	10.19	10.47	10.65	10.53

Female Speech 3

Input SNR	Input LLR	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.76	-2.38	-2.11	-1.46	0.51	1.42
5dB	-1.76	1.90	2.13	2.66	3.87	4.23
10dB	3.24	6.32	6.51	6.93	7.63	7.76
15dB	8.24	10.49	10.65	10.96	11.10	10.75
20dB	13.24	14.66	14.82	15.05	14.94	14.47

Table 6.22: Quality evaluation of various α values for ML estimator based on Gaussian speech model: Segmental SNR (Female speech)

Male Speech 1

Input SNR	Input LLR	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.30	-1.61	-1.31	-0.62	1.45	2.73
5dB	-1.30	2.73	2.98	3.58	5.09	5.84
10dB	3.70	7.08	7.29	7.75	8.57	8.63
15dB	8.70	11.21	11.39	11.74	12.03	11.81
20dB	13.70	15.38	15.50	15.74	15.52	14.97

Male Speech 2

Input SNR	Input LLR	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-5.94	-1.97	-1.72	-1.15	0.34	1.10
5dB	-0.94	2.28	2.49	2.95	3.88	4.17
10dB	4.06	6.59	6.75	7.08	7.45	7.53
15dB	9.06	10.77	10.89	11.10	11.00	10.60
20dB	14.06	15.11	15.21	15.37	15.05	14.60

Male Speech 3

Input SNR	Input LLR	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-7.87	-3.38	-3.09	-2.42	-0.46	0.76
5dB	-2.87	0.84	1.08	1.62	2.91	3.36
10dB	2.13	5.02	5.22	5.63	6.20	6.38
15dB	7.13	9.28	9.43	9.74	9.87	9.55
20dB	12.13	13.42	13.53	13.73	13.48	12.92

Table 6.23: Quality evaluation of various α values for ML estimator based on Gaussian speech model: Segmental SNR (Male speech)

Female Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.02	1.71	1.69	1.67	1.65	1.63
5	1.66	1.30	1.29	1.26	1.33	1.34
10	1.29	0.95	0.94	0.91	1.10	1.16
15	0.91	0.66	0.65	0.63	0.72	0.86
20	0.60	0.45	0.44	0.42	0.50	0.59

Female Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.61	1.45	1.44	1.44	1.51	1.63
5	1.31	1.14	1.13	1.13	1.22	1.33
10	1.06	0.91	0.90	0.90	1.12	1.19
15	0.78	0.65	0.65	0.65	0.82	0.92
20	0.56	0.48	0.48	0.47	0.68	0.70

Female Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.12	1.75	1.73	1.70	1.64	1.66
5	1.72	1.38	1.36	1.33	1.40	1.46
10	1.31	0.95	0.94	0.91	1.02	1.10
15	0.96	0.70	0.70	0.68	0.72	0.87
20	0.64	0.46	0.45	0.43	0.50	0.61

Table 6.24: Quality evaluation of various α values for ML estimator based on Gaussian speech model: Log Likelihood Ratio (Female speech)

Male Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.49	2.04	2.02	1.98	1.98	2.00
5	2.01	1.59	1.57	1.53	1.57	1.47
10	1.58	1.21	1.20	1.17	1.27	1.27
15	1.17	0.89	0.88	0.86	0.94	0.97
20	0.81	0.60	0.60	0.58	0.66	0.75

Male Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.38	1.17	1.17	1.16	1.26	1.32
5	1.08	0.83	0.82	0.80	0.93	1.04
10	0.78	0.63	0.62	0.61	0.73	0.88
15	0.53	0.42	0.41	0.41	0.54	0.74
20	0.36	0.29	0.29	0.29	0.40	0.43

Male Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.67	1.43	1.42	1.41	1.50	1.58
5	1.35	1.12	1.11	1.10	1.29	1.30
10	1.08	0.84	0.84	0.83	0.94	1.03
15	0.79	0.61	0.60	0.60	0.75	0.78
20	0.55	0.42	0.41	0.41	0.50	0.63

Table 6.25: Quality evaluation of various α values for ML estimator based on Gaussian speech model: Log Likelihood Ratio (Male speech)

We now repeat the simulations under the more realistic assumption of estimating the prior SNR ξ from the actual noisy speech. The results are given in Tables 6.20 to 6.25. The variance of the noisy speech component is also estimated with a low-pass filter:

$$\widehat{\lambda_Y}(m) = \beta_Y \widehat{\lambda_Y}(m-1) + (1-\beta_Y)|Y(m)|^2. \tag{6.60}$$

The time constant here is chosen to be 10 ms. And the prior SNR is estimated as:

$$\xi = \max(\frac{\lambda_Y}{\lambda_N} - 1, 10^{-3}). \tag{6.61}$$

The threshold value is chosen to keep the ξ value always positive.

The simulation results show:

• Most of the tests lead to better quality of the enhanced speech with the PESQ increasing, segmental SNR increasing and LLR deceasing. The level of improvement is considerably lower than the results with ideal variance estimation.

• α=1 leads to the best PESQ scores overall. For the three smaller α choices, performance was affected more significantly by the non-ideal variance estimation.

• The segmental SNR results still show the similar trend: larger α leads to better noise removal. At 0 dB input SNRs, α=1 provides 2.3-3.3 dB more segmental SNR improvement than α=0.2 (comparing to 0.15-0.8 dB with ideal speech variance estimation).

• The log-likelihood ratio (LLR) results show that α=0.5 has lower LLRs, though α=0.2 or 0.3 have close results. α=1 or 1.5 still have considerably higher LLRs.

• Informal subjective listening tests show that the enhanced speech removes the additive noise efficiently. For lower SNR cases, some musical noise is introduced to the enhanced speech. α=0.5 or 1 provide similar quality with α=1 having higher distortion and better noise reduction. For higher SNR cases, the distortion is not noticeable and α=1 has lower background noise.

• Unlike the MMSE-based algorithm (Tables 6.1 to 6.6), the ML approach consistently improves the quality for all the αs except α =1.5. The ML approach performs slightly worse than the MMSE approaches with lower input SNRs, but with significantly lower complexity.

6.3.3 ML-Based Speech Enhancement Algorithm with Gaussian Speech Model and Voice Activity Detector

In this section, we study the impact of adding a voice activity detector (VAD) as we did for the MMSE. As for the smaller αs, the proposed ML algorithm amplifies the weak frames when the prior SNR is high. This lead to unpleasant artifacts as for the MMSE case. Thus the same voice activity detector used for the MMSE algorithm is used here. We expect further improvement of quality, especially for the smaller α values.

Tables 6.26 to 6.31 show the results for the ML estimator based on the Gaussian speech model with a voice activity detector. Of the three quality measures, α=0.5 is the best overall and 0.2 and 0.3 are just slightly worse. The difference was not noticeable in informal listening tests. The impact of the VAD is more significant with smaller αs. This is due to the fact that the attenuation at low instantaneous SNR for larger αs is already low.

The PESQ results show further improvement over the algorithm without the VAD. The results are comparable with the MMSE-based estimator with Gaussian speech model. The difference is usually within 0.1. The ML-based approach does provide improvement for the distortion (LLR) measure, but has slightly lower segmental SNR improvement.

When the Voice Activity Detector (VAD) is applied, the quality is further improved. More weak frames are removed rather than amplified without the VAD. Comparing to the MMSE-based algorithm (Tables 6.7 to 6.12), the segmental SNR improvement is lower and

the distortion is also lower due to the fact that ML-based algorithm left the high instantaneous SNR frames unchanged rather than attenuated. Thus, less noise is removed to ensure that the distortion to the speech is kept low.

6.3.4 Performance Analysis of Proposed ML-based Speech Enhancement with Gaussian Speech Model

In this section, equation (6.54) is analyzed for further understanding of the algorithm. First, the equation is rewritten as a gain function G form with respect to the input y.

$$G = \frac{\hat{s}_{ML}}{v} = \frac{\lambda_X + \sqrt{\lambda_X^2 + 2\left(1-2\alpha\right)\lambda_X\lambda_N\left(\lambda_X+\lambda_N\right)/(v^2)}}{2\left(\lambda_X+\lambda_N\right)} = \frac{\xi + \sqrt{\xi^2 + 2(1-2\alpha)\xi(\xi+1)/\gamma}}{2\left(\xi+1\right)} \tag{6.62}$$

where $\xi = \frac{\lambda_X}{\lambda_N}$ and $\gamma = \frac{v^2}{\lambda_N}$ are the *a priori* and *a posteriori* SNR respectively.

The gain G is depicted as functions of ξ and the instantaneous SNR $\gamma - 1$. Three different figures are shown for three α values of 0.3, 0.5 and 1.

Female Speech 1

Input SNR	Input LLR	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.36	+0.39	+0.39	+0.41	+0.44	+0.14
5dB	1.64	+0.43	+0.45	+0.47	+0.49	+0.32
10dB	1.93	+0.49	+0.49	+0.51	+0.50	+0.35
15dB	2.32	+0.46	+0.47	+0.49	+0.49	+0.39
20dB	2.67	+0.44	+0.45	+0.46	+0.47	+0.38

Female Speech 2

Input SNR	Input LLR	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.60	+0.24	+0.25	+0.26	+0.23	-0.01
5dB	1.87	+0.32	+0.32	+0.34	+0.26	-0.19
10dB	2.17	+0.32	+0.32	+0.33	+0.20	-0.06
15dB	2.44	+0.33	+0.33	+0.33	+0.27	+0.00
20dB	2.73	+0.36	+0.36	+0.37	+0.39	+0.23

Female Speech 3

Input SNR	Input LLR	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.25	+0.49	+0.52	+0.54	+0.53	+0.51
5dB	1.60	+0.54	+0.55	+0.58	+0.60	+0.58
10dB	1.96	+0.54	+0.55	+0.57	+0.61	+0.54
15dB	2.33	+0.47	+0.48	+0.50	+0.52	+0.45
20dB	2.70	+0.44	+0.45	+0.48	+0.53	+0.45

Table 6.26: Quality evaluation of various α values for ML estimator based on Gaussian speech model (with VAD): PESQ improvements (Female speech)

Male Speech 1

Input SNR	Input LLR	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.39	+0.29	+0.29	+0.31	+0.32	+0.22
5dB	1.57	+0.40	+0.41	+0.43	+0.43	+0.21
10dB	1.83	+0.47	+0.47	+0.49	+0.49	+0.24
15dB	2.15	+0.41	+0.42	+0.43	+0.39	+0.31
20dB	2.49	+0.33	+0.34	+0.34	+0.29	+0.19

Male Speech 2

Input SNR	Input LLR	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.63	+0.42	+0.42	+0.43	+0.36	+0.13
5dB	1.95	+0.37	+0.38	+0.40	+0.42	+0.32
10dB	2.29	+0.52	+0.54	+0.55	+0.54	+0.44
15dB	2.68	+0.49	+0.50	+0.51	+0.47	+0.34
20dB	3.11	+0.28	+0.28	+0.31	+0.27	+0.25

Male Speech 3

Input SNR	Input LLR	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.35	+0.46	+0.47	+0.49	+0.54	+0.36
5dB	1.70	+0.53	+0.55	+0.57	+0.63	+0.51
10dB	2.06	+0.60	+0.62	+0.63	+0.64	+0.52
15dB	2.47	+0.53	+0.54	+0.56	+0.56	+0.42
20dB	2.84	+0.41	+0.42	+0.44	+0.46	+0.34

Table 6.27: Quality evaluation of various α values for ML estimator based on Gaussian speech model (with VAD): PESQ improvements (Male speech)

Female Speech 1

Input SNR	Input LLR	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.56	0.99	1.17	1.55	2.43	2.71
5dB	-1.56	4.55	4.67	4.93	5.26	5.17
10dB	3.44	8.12	8.20	8.33	8.23	7.78
15dB	8.44	11.67	11.72	11.77	11.37	10.79
20dB	13.44	15.45	15.46	15.46	14.94	14.16

Female Speech 2

Input SNR	Input LLR	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-11.78	-3.87	-3.68	-3.22	-1.85	-0.43
5dB	-6.78	-0.02	0.12	0.48	1.40	2.18
10dB	-1.78	3.55	3.67	3.92	4.40	4.65
15dB	3.22	7.39	7.46	7.62	7.71	7.57
20dB	8.22	11.11	11.16	11.24	11.08	10.73

Female Speech 3

Input SNR	Input LLR	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.76	0.26	0.41	0.73	1.41	1.80
5dB	-1.76	4.16	4.27	4.50	4.87	4.87
10dB	3.24	7.80	7.87	8.02	8.05	7.80
15dB	8.24	11.49	11.54	11.63	11.44	11.00
20dB	13.24	15.36	15.39	15.43	15.10	14.54

Table 6.28: Quality evaluation of various α values for ML estimator based on Gaussian speech model (with VAD): Segmental SNR (Female speech)

Male Speech 1

Input SNR	Input LLR	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.30	1.18	1.35	1.72	2.64	3.05
5dB	-1.30	5.11	5.24	5.51	5.97	5.98
10dB	3.70	8.85	8.94	9.11	9.19	8.91
15dB	8.70	12.42	12.48	12.57	12.36	11.89
20dB	13.70	16.05	16.07	16.09	15.63	14.98

Male Speech 2

Input SNR	Input LLR	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-5.94	0.32	0.46	0.73	1.34	1.53
5dB	-0.94	4.12	4.22	4.40	4.60	4.59
10dB	4.06	7.82	7.89	8.02	7.99	7.69
15dB	9.06	11.55	11.59	11.65	11.46	11.05
20dB	14.06	15.51	15.53	15.54	15.16	14.54

Male Speech 3

Input SNR	Input LLR	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-7.87	-0.72	-0.56	-0.22	0.68	1.33
5dB	-2.87	2.98	3.09	3.33	3.74	3.88
10dB	2.13	6.52	6.60	6.74	6.77	6.51
15dB	7.13	10.27	10.32	10.42	10.21	9.75
20dB	12.13	14.04	14.06	14.09	13.71	13.09

Table 6.29: Quality evaluation of various α values for ML estimator based on Gaussian speech model (with VAD): Segmental SNR (Male speech)

Female Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.02	1.45	1.45	1.44	1.70	1.71
5	1.66	1.09	1.09	1.08	1.25	1.29
10	1.29	0.84	0.84	0.84	1.02	1.16
15	0.91	0.55	0.55	0.56	0.73	0.85
20	0.60	0.37	0.37	0.37	0.50	0.65

Female Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.61	1.24	1.24	1.25	1.51	1.62
5	1.31	1.01	1.00	1.01	1.22	1.38
10	1.06	0.82	0.83	0.84	1.13	1.26
15	0.78	0.54	0.54	0.55	0.81	0.99
20	0.56	0.44	0.44	0.46	0.65	0.75

Female Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.12	1.46	1.45	1.43	1.37	1.53
5	1.72	1.11	1.10	1.09	1.13	1.23
10	1.31	0.80	0.79	0.79	0.96	1.07
15	0.96	0.55	0.55	0.56	0.65	0.76
20	0.64	0.41	0.41	0.41	0.48	0.58

Table 6.30: Quality evaluation of various α values for ML estimator based on Gaussian speech model (with VAD): Log Likelihood Ratio (Female speech)

Male Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.49	1.74	1.72	1.71	1.70	1.69
5	2.01	1.35	1.34	1.33	1.34	1.35
10	1.58	1.01	1.00	1.00	1.15	1.19
15	1.17	0.77	0.77	0.76	0.92	0.97
20	0.81	0.55	0.55	0.56	0.69	0.75

Male Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.38	0.97	0.97	0.96	1.13	1.31
5	1.08	0.72	0.72	0.72	0.92	1.14
10	0.78	0.53	0.53	0.53	0.65	0.78
15	0.53	0.32	0.32	0.32	0.49	0.67
20	0.36	0.27	0.27	0.28	0.35	0.46

Male Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.67	1.13	1.12	1.11	1.30	1.58
5	1.35	0.91	0.91	0.90	1.09	1.30
10	1.08	0.64	0.64	0.63	0.74	0.93
15	0.79	0.51	0.50	0.49	0.65	0.75
20	0.55	0.35	0.35	0.36	0.48	0.60

Table 6.31: Quality evaluation of various α values for ML estimator based on Gaussian speech model (with VAD): Log Likelihood Ratio (Male speech)

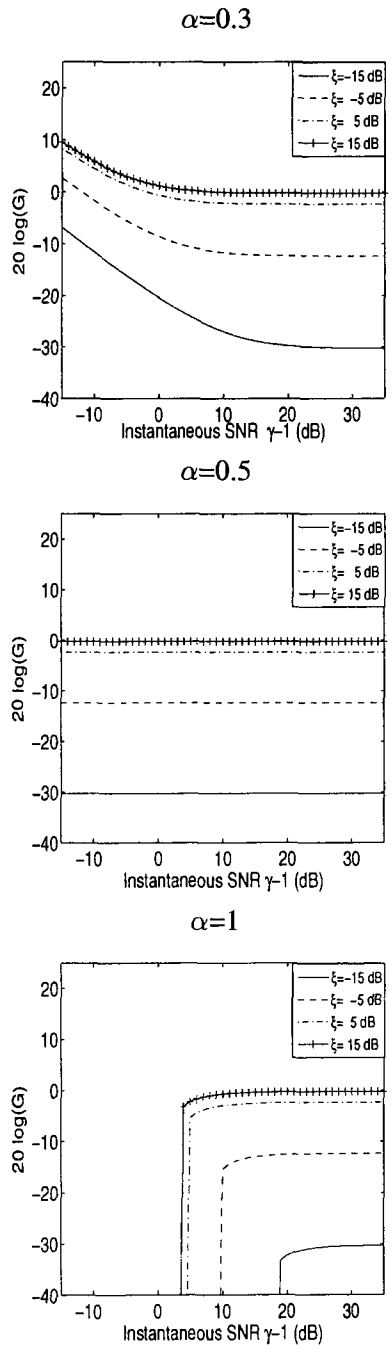

Figure 6.4: The gain function for Maximum Likelihood speech enhancement algorithm with Gaussian speech model

From Figure 6.4, we can see that the gain functions are similar for high γs irrespective of the α values. This means when γ is high, the gain is decided by the prior SNR ξ only. Larger ξ lead to higher gains for the same γ, which means the more noisy the speech is, the more it will be attenuated to remove the noise.

When the prior SNR is fixed, α=0.3 gives higher gains comparing to higher γ; α=0.5 have the same gain for all the γ and α=1 provides lower gains for lower γ. This means α=0.3 will amplify the weak speeches with high ξ and α=1 will attenuate them. This amplification with α=0.3 improves the quality of enhanced speech with less distortion and better PESQ score (Tables 6.14, 6.15, 6.18 and 6.19).

On the other hand, α=0.5 provides better noise removal than smaller αs (Tables 6.16 and 6.17). This can be explained by the lower gains for the same γ values. Speech will be compressed depending on ξ values. Neither speech nor noise will be amplified with α=0.5.

For α=1 or 1.5, more speech/noise is removed from the original noisy signal. It leads to distortion of the speech signal as a result of this aggressive noise removal.

When the noisy signal is used for the estimation of the speech variance, the prior SNR estimate has some error. A typical comparison is shown in Figure 6.5. At the low prior SNR frames, the estimated prior SNRs tend to be higher than their actual value. For the high SNR frames, the estimated prior SNRs tend to be lower than their actual value. Also the lower prior SNRs are, the larger errors in the corresponding prior SNR estimation.

When α=0.3, these errors in prior SNR estimation can lead to amplification of the weak frames in the noisy speech. Under such condition, the speech quality is affected considerably. This effect is shown by the speech quality scores for α=0.3 (so is α=0.2) in Tables 6.20 to 6.25.

With ideal prior SNR estimation, $\alpha = 1$ leads to 0.1-0.35 PESQ degradation compared

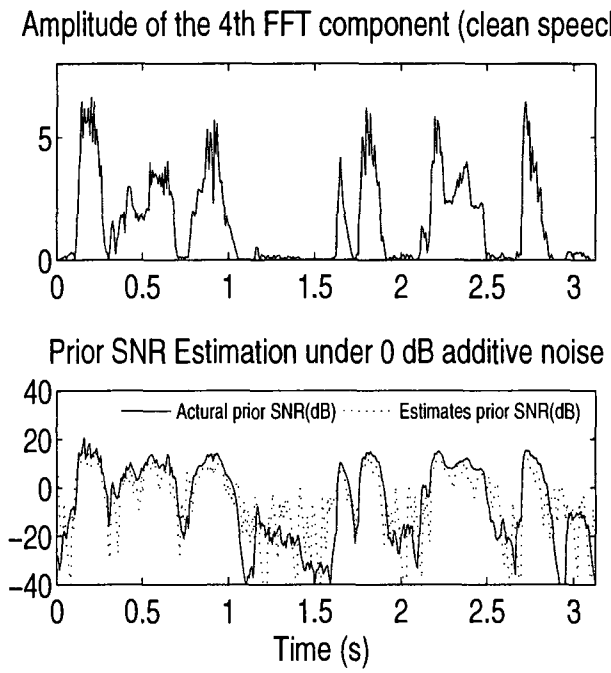

Figure 6.5: A typical comparison between the estimated prior SNR under 0 dB additive noise and the actual prior SNR

to smaller αs. The drawback is that a certain portion of the weak speech is removed together with the noise, which leads to higher distortion. When the instantaneous SNR is low and the prior SNR estimation is less accurate, the algorithm will remove the noise and speech no matter how the prior SNR is estimated. So the results are less affected by the prior SNR estimation. The overall speech quality is better than using smaller αs. Tables 6.20 and 6.25 show that α=1 leads to the best overall PESQ scores. But the performance is still significantly lower than the ideal case.

6.4 Proposed ML-based Speech Enhancement Algorithm with Laplacian Speech Model

6.4.1 The Likelihood Function for the Laplacian Speech Model and its Maximization

In this subsection, we repeat the ML approach in section 6.3 except that we assume a Laplacian distribution for the spectral amplitude of the speech signal:

$$f_X(X) = \frac{1}{a} e^{-\frac{|X|}{a}} = \frac{1}{a} e^{-\frac{s}{a}}. \tag{6.63}$$

where $s = |X|$ is amplitude of the spectrum of the clean signal and a is the factor of the Laplacian distribution. Let $t = s^{2\alpha}$ represent the loudness measure of X. In ML estimation, we want to maximize the probability of $f(t|Y)$, where Y is the spectrum of the noisy signal. The PDF of t can be derived as following:

$$f_T(t) = \frac{f_X(t^{\frac{1}{2\alpha}})}{\left|\frac{d(x^{2\alpha})}{dx}\right|_{x=t^{\frac{1}{2\alpha}}}} = \frac{f_X(t^{\frac{1}{2\alpha}})}{2\alpha t^{\frac{2\alpha-1}{2\alpha}}} = \frac{e^{-\frac{t^{\frac{1}{2\alpha}}}{a}}}{2\alpha t^{\frac{2\alpha-1}{2\alpha}} \cdot a}. \tag{6.64}$$

The conditional probability function of t is given by:

$$f(t|Y) = f(t, Y)/f(Y) \tag{6.65}$$

Maximizing $f(t|Y)$ given Y is equivalent to maximizing $f(t, Y)$. As t is a function of X only, we can assume that t is independent with the additive noise spectrum N. Recalling that $N = Y - X$, then:

$$f_{t,Y}(t, Y) = f_t(t) f_N(Y - X) = \frac{e^{-\frac{t^{\frac{1}{2\alpha}}}{a}}}{2\alpha t^{\frac{2\alpha-1}{2\alpha}} \cdot a} \cdot \frac{1}{\pi \lambda_N} e^{-\frac{|Y-X|^2}{\lambda_N}} \tag{6.66}$$

Similar with the discussion for ML-based algorithm with Gaussian speech model, when Y and t are fixed, $f_{t,Y}(t,Y)$ can be maximized by maximizing $f_N(Y-X)$, i.e., minimizing $|Y-X|$. It can be achieved when and only when X has the same phase with Y. Thus

$$|Y-X| = v - t^{\frac{1}{2\alpha}} \tag{6.67}$$

Again differentiating (6.66) and setting the result to 0, we get:

$$C_L \cdot \frac{t^{\frac{2\alpha-1}{2\alpha}} \left(-\frac{\frac{1}{2\alpha} t^{\left(\frac{1}{2\alpha}-1\right)}}{a} - \frac{\frac{1}{\alpha} t^{\left(\frac{1}{\alpha}-1\right)} - 2v\frac{1}{2\alpha} t^{\left(\frac{1}{2\alpha}-1\right)}}{\lambda_N} \right) - \left(\frac{2\alpha-1}{2\alpha}\right) t^{-\frac{1}{2\alpha}}}{\left(t^{\frac{2\alpha-1}{2\alpha}}\right)^2} \cdot e^{-\frac{t^{\frac{1}{2\alpha}}}{a} - \frac{\left(v-t^{\frac{1}{2\alpha}}\right)^2}{\lambda_N}} = 0 \tag{6.68}$$

where $C_L = \frac{1}{2\alpha a \pi \lambda_N}$.

Noting that $e^{-\frac{t^{\frac{1}{2\alpha}}}{a} - \frac{\left(v-t^{\frac{1}{2\alpha}}\right)^2}{\lambda_N}} > 0$ and $t > 0$:

$$-\frac{1}{\alpha\lambda_N} t^{\frac{1}{\alpha}} - \frac{(\lambda_N - 2av)}{2a\alpha\lambda_N} t^{\frac{1}{2\alpha}} - \left(\frac{2\alpha-1}{2\alpha}\right) = 0, \tag{6.69}$$

Multiplying by $a\alpha\lambda_N$, we get:

$$2at^{\frac{1}{\alpha}} + (\lambda_N - 2av)\, t^{\frac{1}{2\alpha}} + (2\alpha-1)\, a\lambda_N = 0. \tag{6.70}$$

Let $\hat{s} = t^{\frac{1}{2\alpha}}$, then:

$$2a\hat{s}^2 + (\lambda_N - 2av)\, \hat{s} + (2\alpha-1)\, a\lambda_N = 0, \tag{6.71}$$

This is a quadratic equation with the discriminant:

$$\Delta_L = (2av - \lambda_N)^2 + 8(1-2\alpha)\, a^2\lambda_N \tag{6.72}$$

When $\Delta_L < 0$, (6.69) is negative for all $t > 0$. So the first derivative of (6.66) is always negative, which means (6.66) is strictly decreasing. It is maximized when $t = 0$, i.e., $\hat{s} = 0$.

When $\Delta_L \geq 0$, the solution is

$$\hat{s}_{L1,L2} = \frac{2av - \lambda_N \pm \sqrt{(2av - \lambda_N)^2 + 8(1-2\alpha)a^2\lambda_N}}{4a} \tag{6.73}$$

Similar to the case for the Gaussian speech model, the second derivative of (6.66) is needed to decide if $f_{t,Y}(t,Y)$ is maximized or minimized at $\hat{s}_{L1,L2}$.

The first derivative is rewritten as

$$\frac{d}{dt} f_{t,Y}(t,Y) = C' F_{L1}(t) F_{L2}(t) \tag{6.74}$$

where

$$C'_L = C_L \frac{1}{2a\alpha\lambda_N} \tag{6.75}$$

$$F_{L1}(t) = 2a \frac{t^{\frac{-1}{2\alpha}}}{-\left(t^{\frac{2\alpha-1}{2\alpha}}\right)^2} \cdot e^{-\frac{t^{\frac{1}{2\alpha}}}{a} - \frac{(v - t^{\frac{1}{\alpha}})^2}{\lambda_N}} \tag{6.76}$$

and

$$F_{L2}(t) = -2at^{\frac{1}{\alpha}} - (\lambda_N - 2av)\, t^{\frac{1}{2\alpha}} - (2\alpha - 1)\, a\lambda_N \tag{6.77}$$

The second derivative of $f_{t,Y}(t,Y)$ at $t = \hat{s}^{2\alpha}_{L1,L2}$ is

$$\begin{aligned} \frac{d^2}{dt^2} f_{t,Y}(t,Y)\Big|_{t=\hat{s}^{2\alpha}_{L1,L2}} &= \frac{d}{dt}\left[C' F_{L1}(t) F_{L2}(t)\right]\Big|_{t=\hat{s}^{2\alpha}_{L1,L2}} \\ &= \left[C' F_{L1}(t)\frac{d}{dt}F_{L2}(t) + C' F_{L1}(t)\frac{d}{dt}F_{L2}(t)\right]\Big|_{t=\hat{s}^{2\alpha}_{L1,L2}} . \end{aligned} \tag{6.78}$$

As

$$F_{L2}(t)|_{t=\hat{s}^{2\alpha}_{L1,L2}} = 0, \tag{6.79}$$

$$\frac{d^2}{dt^2} f_{t,Y}(t,Y)\Big|_{t=\hat{s}^{2\alpha}_{L1,L2}} = \left[C' F_{L1}(t)\frac{d}{dt}F_{L2}(t)\right]\Big|_{t=\hat{s}^{2\alpha}_{L1,L2}} . \tag{6.80}$$

As we have $C_L' \geq 0$, $F_1(t) \geq 0$.

$$\begin{aligned}\frac{d}{dt}F_2(t)\Big|_{t=\hat{s}_{L1,L2}^{2\alpha}} &= \left[\frac{1}{\alpha}(-2a)\,t^{\frac{1}{\alpha}-1} - \frac{1}{2\alpha}(\lambda_N - 2av)\,t^{\frac{1}{2\alpha}-1}\right]\Bigg|_{t=\hat{s}_{L1,L2}^{2\alpha}} \\ &= \left[\frac{(-4a)\,t^{\frac{1}{\alpha}} - (\lambda_N - 2av)\,t^{\frac{1}{2\alpha}}}{2\alpha t}\right]\Bigg|_{t=\hat{s}_{L1,L2}^{2\alpha}} \\ &= \left[\frac{(-4a)\,s^2 - (\lambda_N - 2av)\,s}{2\alpha s^{2\alpha}}\right]\Bigg|_{s=\hat{s}_{L1,L2}} \end{aligned} \tag{6.81}$$

Assigning $\hat{s}_1$ to the root with the $+$ sign, then

$$\begin{aligned}\frac{d}{dt}F_2(t)\Big|_{t=\hat{s}_{L1}^{2\alpha}} &= \frac{(-4a)\left(\frac{(2av-\lambda_N)+\sqrt{\Delta_L}}{4a}\right)^2 - (\lambda_N - 2av)\left(\frac{(2av-\lambda_N)+\sqrt{\Delta_L}}{(4a)}\right)}{2\alpha s^{2\alpha}} \\ &= \frac{-\left((2av-\lambda_N)+\sqrt{\Delta_L}\right)^2 - (\lambda_N - 2av)\left(2av-\lambda_N+\sqrt{\Delta_L}\right)}{8a\alpha s^{2\alpha}} \\ &= \frac{-(2av-\lambda_N)\sqrt{\Delta_L} - \Delta_L}{8a\alpha s_{L1}^{2\alpha}} \\ &= \frac{-s_{L1}\sqrt{\Delta_L}}{8a\alpha s_{L1}^{2\alpha}} \end{aligned} \tag{6.82}$$

thus as long as $\hat{s}_{L1} \geq 0$, it corresponds to the maximum of $f_{t,Y}(t, Y)$ under the Laplacian assumption. If $\hat{s}_{L1} < 0$, both roots are negative and (6.66) is strictly decreasing when $t \geq 0$. So the maximum is reached when $t = 0$.

Also,

$$\begin{aligned}\frac{d}{dt}F_2(t)\Big|_{t=\hat{s}_{L2}^{2\alpha}} &= \frac{\left((2av-\lambda_N)+\sqrt{\Delta_L}\right)^2 - (\lambda_N - 2av)\left(2av-\lambda_N+\sqrt{\Delta_L}\right)}{8a\alpha s^{2\alpha}} \\ &= \frac{-(2av-\lambda_N)\sqrt{\Delta_L} - \Delta_L}{8a\alpha s_{L2}^{2\alpha}} \\ &= \frac{s_{L2}\sqrt{\Delta_L}}{8a\alpha s_{L2}^{2\alpha}} \end{aligned} \tag{6.83}$$

There are two possibilities, if $s_{L2} \geq 0$, then the above

$$\frac{d}{dt}F_2(t)\Big|_{t=\hat{s}_{L2}^{2\alpha}} \geq 0 \tag{6.84}$$

which means $\hat{s}_2$ corresponds to the minimum of $f_{t,Y}(t, Y)$. Note that $s_{L2} < 0$ must be discarded since s is an estimation of amplitude and can't be negative. Either way, $\hat{s}_{L1}$ corresponds to the only maximum of $f_{t,Y}(t, Y)$.

Overall, the maximum likelihood under the Laplacian assumption is

$$\hat{s}_{ML,L} = \max\left(\frac{2av - \lambda_N + \sqrt{\Delta_L}}{4a}, 0\right) \tag{6.85}$$

when

$$\Delta_L = (2av - \lambda_N)^2 + 8(1 - 2\alpha)a^2\lambda_N \geq 0 \tag{6.86}$$

and for $\Delta_L < 0$

$$\hat{s}_{ML,L} = 0. \tag{6.87}$$

6.4.2 Simulation Results for Proposed ML-based Algorithm with Laplacian Speech Model

In this section, the proposed speech enhancement algorithm based on the ML estimation is simulated. The speech signal amplitude components in the DFT domain are enhanced using the proposed ML estimation with Laplacian assumptions as in (6.73). The complete speech enhancement algorithm is shown in Table 6.4.2

The estimation of the variance of noise and noisy speech also follow (6.57) and (6.58). Then, a is estimated as $\sqrt{\max(\lambda_Y - \lambda_N, 0)/2}$. The results for ideal variance estimation are shown. In this case, we estimate the speech signal variance λ_X using the clean speech as in (6.59). Under such assumption, the performance of the proposed approach are shown in Tables 6.33 to 6.38 different α values: 0.2, 0.3, 0.5, 0.7 and 1.

β_Y is chosen to let the time constant of the adaptive process to be 10 ms.

β_N is chosen to let the timeconstant of the adaptive process to be 0.5 sec.

For each time step m do,

$$\bar{Y}(m) = FFT\left(\bar{y}(m)\right)$$
$$\bar{n}(m) \leftarrow \text{Read from noise memory}$$
$$\bar{N}(m) = FFT\left(\bar{n}(m)\right)$$
$$\hat{V}(m) = [v_1(m)|v_2(m)|\ldots|v_R(m)] = \left|\bar{Y}(m)\right|;$$
$$\hat{\theta}(m) = [\theta_1(m)|\theta_2(m)|\ldots|\theta_R(m)] = \angle\bar{Y}(m);$$
$$\hat{U}(m) = [u_1(m)|u_2(m)|\ldots|u_R(m)] = \left|\bar{N}(m)\right|;$$

For i= 1, 2,. . ., R do

$$\lambda_{Yi}(m) = \beta_Y\lambda_{Yi}(m-1) + (1-\beta_Y)v_i^2(m)$$
$$\lambda_{Ni}(m) = \beta_N\sigma_i^2(m-1) + (1-\beta_N)u_i^2(m)$$
$$a_i(m) = \max\left(\sqrt{\frac{\lambda_{Yi}(m)-\lambda_{Ni}(m)}{2}}, 0\right)$$
$$\Delta_{Li}(m) = \left(2a_i(m)v_i(m) - \lambda_{Ni}(m)\right)^2 + 8\left(1-2\alpha\right)a_i^2(m)\lambda_{Ni}(m)$$

When $\Delta_{Li}(m) \geq 0$

$$\hat{s}_i(m) = \max\left(\frac{2a_i(m)v_i(m) - \lambda_{Ni}(m) + \sqrt{\Delta_{Li}(m)}}{4a_i(m)}, 0\right)$$

When $\Delta_{Li}(m) < 0$

$$\hat{s}_i(m) = 0$$

end;

$$\hat{S}(m) = [\hat{s}_1(m)|\hat{s}_2(m)|\ldots|\hat{s}_R(m)]$$
$$\hat{X}(m) = \hat{S}(m)\cdot e^{j\theta(m)}$$
$$\hat{x}(m) = IFFT\left(\hat{X}(m)\right)$$

Table 6.32: Proposed ML-based Speech Enhancement Algorithm (Laplacian Speech Model)

Female Speech 1

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0dB	1.36	+1.00	+1.03	+0.57	+0.42	+0.26
5dB	1.64	+0.84	+0.88	+0.72	+0.37	+0.14
10dB	1.93	+0.73	+0.79	+0.74	+0.56	+0.35
15dB	2.32	+0.60	+0.66	+0.66	+0.60	+0.48
20dB	2.67	+0.48	+0.54	+0.61	+0.58	+0.49

Female Speech 2

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0dB	1.60	+0.81	+0.81	+0.19	-0.01	-0.14
5dB	1.87	+0.70	+0.71	+0.29	-0.05	-0.28
10dB	2.17	+0.65	+0.67	+0.52	+0.34	+0.01
15dB	2.44	+0.55	+0.57	+0.50	+0.40	+0.30
20dB	2.73	+0.47	+0.50	+0.47	+0.39	+0.34

Female Speech 3

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0dB	1.25	+1.09	+1.10	+0.80	+0.57	+0.25
5dB	1.60	+0.92	+0.94	+0.83	+0.65	+0.48
10dB	1.96	+0.73	+0.76	+0.65	+0.56	+0.48
15dB	2.33	+0.62	+0.67	+0.74	+0.69	+0.50
20dB	2.70	+0.50	+0.54	+0.59	+0.57	+0.41

Table 6.33: Quality evaluation of enhanced speech for various α values for ML estimator based on Laplacian speech model with ideal variance estimation: PESQ improvements (Female speech)

Male Speech 1

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0dB	1.39	+0.99	+1.00	+0.84	+0.65	+0.63
5dB	1.57	+0.89	+0.91	+0.86	+0.66	+0.63
10dB	1.83	+0.79	+0.81	+0.80	+0.63	+0.60
15dB	2.15	+0.61	+0.64	+0.67	+0.57	+0.55
20dB	2.49	+0.44	+0.47	+0.51	+0.47	+0.45

Male Speech 2

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0dB	1.63	+0.90	+0.92	+0.55	+0.38	+0.18
5dB	1.95	+0.80	+0.83	+0.69	+0.35	+0.31
10dB	2.29	+0.71	+0.75	+0.68	+0.50	+0.41
15dB	2.68	+0.60	+0.64	+0.60	+0.42	+0.38
20dB	3.11	+0.38	+0.45	+0.46	+0.37	+0.22

Male Speech 3

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0dB	1.35	+1.22	+1.26	+0.98	+0.62	+0.34
5dB	1.70	+1.09	+1.14	+0.91	+0.81	+0.53
10dB	2.06	+0.86	+0.92	+0.79	+0.63	+0.54
15dB	2.47	+0.67	+0.73	+0.74	+0.60	+0.49
20dB	2.84	+0.46	+0.50	+0.57	+0.50	+0.43

Table 6.34: Quality evaluation of enhanced speech for various α values for ML estimator based on Laplacian speech model with ideal variance estimation: PESQ improvements (Male speech)

Female Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0dB	-6.56	3.53	4.04	4.52	4.28	3.97
5dB	-1.56	5.97	6.42	6.83	6.48	6.01
10dB	3.44	8.73	9.12	9.50	9.16	8.64
15dB	8.44	11.93	12.26	12.56	12.23	11.70
20dB	13.44	15.58	15.82	16.08	15.76	15.20

Female Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0dB	-11.78	2.16	2.78	3.29	3.08	2.89
5dB	-6.78	3.64	4.19	4.74	4.65	4.60
10dB	-1.78	5.96	6.46	6.99	6.79	6.45
15dB	3.22	8.57	8.98	9.51	9.33	9.04
20dB	8.22	11.87	12.19	12.58	12.34	11.89

Female Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0dB	-6.76	3.39	3.90	4.44	4.17	3.87
5dB	-1.76	5.83	6.29	6.84	6.66	6.32
10dB	3.24	8.65	9.05	9.56	9.35	8.97
15dB	8.24	11.92	12.25	12.66	12.44	12.02
20dB	13.24	15.57	15.82	16.13	15.95	15.49

Table 6.35: Quality evaluation of enhanced speech for various α values for ML estimator based on Laplacian speech model with ideal variance estimation: Segmental SNR (Female speech)

Male Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0dB	-6.30	4.48	4.95	5.43	5.26	4.84
5dB	-1.30	7.01	7.46	7.93	7.69	7.30
10dB	3.70	9.85	10.27	10.63	10.32	9.89
15dB	8.70	12.68	13.03	13.43	13.14	12.71
20dB	13.70	16.19	16.46	16.77	16.45	15.90

Male Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0dB	-5.94	3.22	3.70	4.19	4.03	3.74
5dB	-0.94	5.71	6.15	6.65	6.51	6.13
10dB	4.06	8.62	8.99	9.40	9.20	8.89
15dB	9.06	11.94	12.22	12.56	12.31	11.90
20dB	14.06	15.83	16.04	16.28	16.01	15.55

Male Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0dB	-7.87	2.58	3.15	3.82	3.78	3.54
5dB	-2.87	4.95	5.43	5.95	5.76	5.39
10dB	2.13	7.69	8.10	8.51	8.22	7.84
15dB	7.13	10.79	11.12	11.50	11.20	10.71
20dB	12.13	14.46	14.71	15.00	14.71	14.17

Table 6.36: Quality evaluation of enhanced speech for various α values for ML estimator based on Laplacian speech model with ideal variance estimation: Segmental SNR (Male speech)

Female Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0	2.02	0.44	0.46	0.92	1.01	1.06
5	1.66	0.41	0.41	0.86	0.96	0.98
10	1.29	0.41	0.40	0.68	0.80	0.86
15	0.91	0.34	0.33	0.45	0.57	0.65
20	0.60	0.28	0.27	0.33	0.40	0.47

Female Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0	1.61	0.43	0.44	0.86	0.98	0.99
5	1.31	0.43	0.44	0.79	0.88	0.88
10	1.06	0.38	0.39	0.74	0.90	0.95
15	0.78	0.33	0.33	0.51	0.57	0.66
20	0.56	0.28	0.28	0.42	0.43	0.50

Female Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0	2.12	0.47	0.48	0.91	1.03	1.00
5	1.72	0.44	0.44	0.78	0.88	0.94
10	1.31	0.43	0.41	0.51	0.68	0.85
15	0.96	0.39	0.38	0.43	0.54	0.61
20	0.64	0.32	0.31	0.40	0.42	0.47

Table 6.37: Quality evaluation of enhanced speech for various α values for ML estimator based on Laplacian speech model with ideal variance estimation: Log Likelihood Ratio (Female speech)

Male Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0	2.49	0.61	0.61	0.95	1.02	1.04
5	2.01	0.61	0.59	0.86	1.02	1.03
10	1.58	0.58	0.56	0.72	0.85	0.91
15	1.17	0.55	0.53	0.67	0.73	0.79
20	0.81	0.42	0.41	0.50	0.55	0.63

Male Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0	1.38	0.39	0.40	0.76	0.85	0.90
5	1.08	0.36	0.36	0.60	0.67	0.73
10	0.78	0.34	0.33	0.56	0.62	0.76
15	0.53	0.29	0.29	0.31	0.42	0.51
20	0.36	0.25	0.25	0.31	0.31	0.36

Male Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=0.7	α=1
0	1.67	0.50	0.52	0.98	1.14	1.32
5	1.35	0.44	0.45	0.75	0.87	0.95
10	1.08	0.39	0.38	0.58	0.75	0.82
15	0.79	0.36	0.35	0.46	0.50	0.66
20	0.55	0.31	0.30	0.38	0.37	0.40

Table 6.38: Quality evaluation of enhanced speech for various α values for ML estimator based on Laplacian speech model with ideal variance estimation: Log Likelihood Ratio (Male speech)

From Tables 6.33 to 6.38, we clearly see that:

• All the tested α values provide improved segmental SNR and LLR. Almost all (but one speech with α=0.7 or 1 at 0 or 5 dB input SNR) tests provide improvements in PESQ scores. Overall, the speech quality is improved.

• Of the five selected α values, $\alpha = 0.3$ provides the best PESQ improvements when the input SNR is low. When the input SNR is high, the three α values of 0.2, 0.3 and 0.5 lead to similar PESQ improvements. α=0.5 performs significantly worse for low input SNR cases. The PESQ decreases noticeably when α is 1 or 1.5.

• For the segmental SNRs for the enhanced signals, α=0.5 provides the best segmental SNR improvements. α=0.3 has slightly worse segmental SNR scores. For all five tested α values, the segmental SNR is within a close range comparing to the overall improvements.

• For the distortion measure LLR, α=0.2 and 0.3 provide comparable results. The LLR increases considerably as α increases beyond 0.3. α=1.5 can lead to more distortion overall as measured by LLR.

• Based on informal subjective listening tests, the residual noise level is higher for smaller αs, but the enhanced speech is more distorted for larger αs. α=0.3 provides the best overall quality.

• Comparing with the results under Gaussian assumptions, α=0.3 performs best under both assumptions. The three quality measures all favor the Gaussian assumption. The listening tests show that the Laplacian assumption leads to higher noise residue but less noticeable distortion to the speech signals. This will be analyzed later in section 6.5.1.

Next, the simulations are repeated for actual speech variance estimation from the noisy signal. The noise variance is subtracted from the noisy speech variance as the estimate of the speech variance. The α values tested are chosen to be 0.2, 0.3, 0.5, 1 and 1.5.

Female Speech 1

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.36	+0.20	+0.22	+0.26	+0.31	+0.04
5dB	1.64	+0.23	+0.26	+0.32	+0.41	+0.16
10dB	1.93	+0.29	+0.31	+0.37	+0.43	+0.20
15dB	2.32	+0.24	+0.27	+0.33	+0.41	+0.34
20dB	2.67	+0.22	+0.24	+0.30	+0.39	+0.31

Female Speech 2

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.60	+0.13	+0.14	+0.15	+0.08	-0.40
5dB	1.87	+0.17	+0.19	+0.22	+0.19	-0.22
10dB	2.17	+0.17	+0.19	+0.23	+0.20	-0.05
15dB	2.44	+0.21	+0.23	+0.26	+0.28	+0.13
20dB	2.73	+0.23	+0.25	+0.30	+0.34	+0.19

Female Speech 3

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.25	+0.27	+0.30	+0.36	+0.46	+0.24
5dB	1.60	+0.30	+0.33	+0.41	+0.53	+0.44
10dB	1.96	+0.29	+0.32	+0.39	+0.50	+0.32
15dB	2.33	+0.27	+0.30	+0.36	+0.47	+0.35
20dB	2.70	+0.24	+0.27	+0.33	+0.42	+0.35

Table 6.39: Quality evaluation of the enhanced speech for various α values for ML estimator based on Laplacian speech model: PESQ improvements (Female speech)

Male Speech 1

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.39	+0.16	+0.17	+0.19	+0.19	-0.02
5dB	1.57	+0.20	+0.22	+0.37	+0.38	+0.09
10dB	1.83	+0.27	+0.30	+0.37	+0.35	+0.14
15dB	2.15	+0.21	+0.23	+0.28	+0.30	+0.12
20dB	2.49	+0.21	+0.24	+0.28	+0.26	+0.15

Male Speech 2

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.63	+0.19	+0.21	+0.24	+0.26	+0.15
5dB	1.95	+0.26	+0.29	+0.33	+0.27	+0.07
10dB	2.29	+0.26	+0.28	+0.33	+0.30	+0.10
15dB	2.68	+0.28	+0.31	+0.36	+0.33	+0.28
20dB	3.11	+0.22	+0.25	+0.30	+0.29	+0.21

Male Speech 3

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.35	+0.27	+0.30	+0.36	+0.44	+0.16
5dB	1.70	+0.29	+0.32	+0.42	+0.56	+0.40
10dB	2.06	+0.31	+0.35	+0.43	+0.55	+0.37
15dB	2.47	+0.31	+0.35	+0.41	+0.51	+0.37
20dB	2.84	+0.25	+0.28	+0.34	+0.40	+0.31

Table 6.40: Quality evaluation of the enhanced speech for various α values for ML estimator based on Laplacian speech model: PESQ improvements (Male speech)

Female Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.56	-2.49	-2.01	-0.90	1.41	2.34
5dB	-1.56	2.05	2.47	3.35	4.85	5.10
10dB	3.44	6.35	6.66	7.22	7.76	7.58
15dB	8.44	10.55	10.80	11.09	11.12	10.67
20dB	13.44	14.83	14.97	15.09	14.67	14.01

Female Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-11.78	-7.53	-7.02	-5.82	-2.76	-0.26
5dB	-6.78	-3.16	-2.73	-1.80	0.39	1.92
10dB	-1.78	1.30	1.66	2.38	3.87	4.76
15dB	3.22	5.62	5.88	6.35	7.19	7.52
20dB	8.22	9.96	10.15	10.46	10.73	10.62

Female Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.76	-2.82	-2.35	-1.33	0.72	1.51
5dB	-1.76	1.72	2.10	2.88	4.21	4.77
10dB	3.24	6.04	6.34	6.87	7.54	7.68
15dB	8.24	10.34	10.56	10.88	11.06	10.76
20dB	13.24	14.73	14.88	15.04	14.84	14.37

Table 6.41: Quality evaluation of the enhanced speech for various α values for ML estimator based on Laplacian speech model: Segmental SNR (Female speech)

Male Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.30	-2.08	-1.59	-0.47	1.82	2.66
5dB	-1.30	2.37	2.79	3.66	5.32	5.86
10dB	3.70	6.71	7.05	7.72	8.80	8.80
15dB	8.70	10.95	11.19	11.58	11.88	11.66
20dB	13.70	15.29	15.46	15.62	15.34	14.80

Male Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-5.94	-2.28	-1.87	-0.99	0.62	1.39
5dB	-0.94	2.06	2.37	2.95	3.91	4.40
10dB	4.06	6.41	6.65	7.07	7.51	7.37
15dB	9.06	10.75	10.92	11.14	11.12	10.69
20dB	14.06	15.08	15.18	15.22	14.78	14.29

Male Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-7.87	-3.88	-3.41	-2.37	-0.19	1.25
5dB	-2.87	0.54	0.92	1.70	3.04	3.80
10dB	2.13	4.75	5.05	5.57	6.27	6.49
15dB	7.13	9.09	9.33	9.51	9.64	9.39
20dB	12.13	13.43	13.57	13.72	13.41	12.89

Table 6.42: Quality evaluation of the enhanced speech for various α values for ML estimator based on Laplacian speech model: Segmental SNR (Male speech)

Female Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.02	1.74	1.72	1.68	1.72	1.79
5	1.66	1.33	1.30	1.27	1.39	1.42
10	1.29	0.97	0.95	0.93	1.11	1.09
15	0.91	0.68	0.67	0.65	0.76	0.78
20	0.60	0.44	0.43	0.41	0.54	0.67

Female Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.61	1.48	1.48	1.54	1.59	1.62
5	1.31	1.18	1.17	1.17	1.20	1.23
10	1.06	0.91	0.90	0.91	1.02	1.10
15	0.78	0.66	0.67	0.72	0.97	1.00
20	0.56	0.47	0.46	0.48	0.65	0.72

Female Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.12	1.73	1.70	1.62	1.63	1.68
5	1.72	1.34	1.31	1.24	1.33	1.40
10	1.31	1.04	1.02	0.98	1.04	1.04
15	0.96	0.72	0.70	0.68	0.81	0.88
20	0.64	0.51	0.50	0.49	0.53	0.59

Table 6.43: Quality evaluation of the enhanced speech for various α values for ML estimator based on Laplacian speech model: Log Likelihood Ratio (Female speech)

Male Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.49	2.16	2.12	2.06	2.03	2.04
5	2.01	1.62	1.59	1.51	1.56	1.57
10	1.58	1.23	1.21	1.18	1.25	1.27
15	1.17	0.93	0.91	0.89	0.97	0.98
20	0.81	0.63	0.62	0.59	0.71	0.74

Male Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.38	1.18	1.17	1.16	1.26	1.27
5	1.08	0.87	0.85	0.84	1.07	1.18
10	0.78	0.61	0.59	0.58	0.80	0.92
15	0.53	0.44	0.43	0.43	0.55	0.69
20	0.36	0.29	0.29	0.28	0.36	0.42

Male Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.67	1.41	1.40	1.43	1.70	1.74
5	1.35	1.14	1.12	1.12	1.34	1.37
10	1.08	0.82	0.80	0.80	0.96	1.06
15	0.79	0.64	0.61	0.60	0.78	0.93
20	0.55	0.44	0.43	0.43	0.58	0.61

Table 6.44: Quality evaluation of the enhanced speech for various α values for ML estimator based on Laplacian speech model: Log Likelihood Ratio (Male speech)

Tables 6.39 to 6.44 show the test results using speech variance estimation from the actual noisy speech. The results show:

- Most of the tests still lead to better quality of the enhanced speech with the PESQ increasing, segmental SNR increasing and LLR deceasing. However, the level of improvements is considerably lower than the results with ideal variance estimation.
- α=1 and α=0.5 lead to the best PESQ scores overall. The three smaller α choices were affected more significantly with the non-ideal variance estimation.
- The segmental SNR results show the same trend: larger α leads to better noise removal. At 0 dB input SNRs, α=1 provides 2.7-4.8 dB more sngSNR improvement than α=0.2 (comparing to 0.8-1.2 dB with ideal speech variance estimation). For high input SNRs, the segmental SNR peaks at α=0.5 for 20dB and α=1 for 15 dB. Higher α values leads to more over-attenuation of the original speech.
- The LLR results show that α=0.2, 0.3 or 0.5 results in similar performance. α=1 or 1.5 still has considerably higher LLRs. The overall segmental SNR improvement for α=1 may compensate for the excess distortion and still leads to better overall quality (PESQ scores)
- The informal subjective listening test shows that the enhancement algorithm removes the additive noise efficiently. For lower SNR cases, some musical noise is introduced to the enhanced speech. α=0.5 or 1 provides similar quality and α=1 leads to higher distortion and better noise reduction.
- Overall the PESQ scores are similar to Gaussian assumptions for α=1. The Laplacian assumption leads to higher background noise than the Gaussian assumption. The Laplacian assumption also provides lower distortion to the speech based on the informal subjective listening tests.

6.5 ML-Based Speech Enhancement Algorithm with Laplacian Speech Model and Voice Activity Detector

In this section, a voice activity detector is applied in the speech enhancement algorithm. The VAD used is based on the Laplacian-Gaussian mixture model for speech-noise [46] [118]. This algorithm is designed on the assumption that the noisy speech is a Laplacian-Gaussian mixture (Laplacian for speech and Gaussian for noise).

The simulation results is given in the following Tables 6.45 to 6.50.

From Tables 6.45 to 6.50, we can see that the enhanced speech quality is further improved as measured by most objective measures for smaller αs (0.2 and 0.3). This is due to the attenuation of the weak frames rather than amplification without the VAD. The PESQ score is improved by 0.1 to 0.2 for these αs. The segmental SNR can be improved up to 2.2 dB. Also the LLR distortion measure is lowered.

For the PESQ scores, α=1 still provide the best improvements, though the smaller α's result is only slightly lower. Also smaller αs provides lower distortion measured by LLR. The segmental SNR results are comparable for all the α values with high input SNRs. For the lower input SNRs, larger α still performs better due to the fact that more noise is reduced from the original speech.

Female Speech 1

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.36	+0.27	+0.28	+0.30	+0.32	+0.18
5dB	1.64	+0.34	+0.36	+0.39	+0.35	+0.06
10dB	1.93	+0.40	+0.42	+0.44	+0.46	+0.26
15dB	2.32	+0.34	+0.36	+0.38	+0.41	+0.24
20dB	2.67	+0.37	+0.39	+0.41	+0.42	+0.35

Female Speech 2

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.60	+0.18	+0.19	+0.18	+0.10	-0.30
5dB	1.87	+0.24	+0.24	+0.24	+0.14	-0.23
10dB	2.17	+0.23	+0.24	+0.25	+0.17	-0.20
15dB	2.44	+0.31	+0.32	+0.33	+0.28	+0.03
20dB	2.73	+0.29	+0.30	+0.34	+0.37	+0.23

Female Speech 3

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.25	+0.38	+0.39	+0.41	+0.45	+0.13
5dB	1.60	+0.42	+0.43	+0.44	+0.50	+0.23
10dB	1.96	+0.43	+0.45	+0.47	+0.51	+0.42
15dB	2.33	+0.40	+0.41	+0.43	+0.48	+0.35
20dB	2.70	+0.37	+0.39	+0.41	+0.44	+0.35

Table 6.45: Quality evaluation of the enhanced speech for various α values for ML estimator based on Laplacian speech model (with VAD): PESQ improvements (Female speech)

Male Speech 1

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.39	+0.18	+0.19	+0.21	+0.21	-0.14
5dB	1.57	+0.27	+0.28	+0.31	+0.33	+0.09
10dB	1.83	+0.34	+0.35	+0.38	+0.39	+0.29
15dB	2.15	+0.33	+0.35	+0.36	+0.30	+0.18
20dB	2.49	+0.29	+0.30	+0.30	+0.26	+0.14

Male Speech 2

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.63	+0.29	+0.39	+0.29	+0.27	+0.25
5dB	1.95	+0.36	+0.39	+0.33	+0.26	+0.27
10dB	2.29	+0.53	+0.55	+0.43	+0.38	+0.30
15dB	2.68	+0.38	+0.46	+0.46	+0.39	+0.28
20dB	3.11	+0.24	+0.28	+0.30	+0.26	+0.21

Male Speech 3

Input SNR	Input PESQ	PESQ score improvements				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	1.35	+0.36	+0.38	+0.41	+0.43	+0.28
5dB	1.70	+0.44	+0.45	+0.50	+0.56	+0.28
10dB	2.06	+0.46	+0.48	+0.51	+0.55	+0.32
15dB	2.47	+0.43	+0.44	+0.47	+0.46	+0.33
20dB	2.84	+0.47	+0.48	+0.51	+0.49	+0.36

Table 6.46: Quality evaluation of the enhanced speech for various α values for ML estimator based on Laplacian speech model (with VAD): PESQ improvements (Male speech)

Female Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.56	-0.35	-0.08	0.48	1.81	2.65
5dB	-1.56	3.65	3.84	4.20	4.83	4.96
10dB	3.44	7.44	7.56	7.72	7.86	7.67
15dB	8.44	11.25	11.30	11.32	11.08	10.59
20dB	13.44	15.23	15.27	15.19	14.71	14.06

Female Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-11.78	-5.42	-5.12	-4.40	-2.37	-0.22
5dB	-6.78	-1.24	-0.99	-0.48	0.84	2.01
10dB	-1.78	2.86	3.03	3.34	4.06	4.77
15dB	3.22	6.69	6.80	6.95	7.24	7.31
20dB	8.22	10.74	10.81	10.88	10.83	10.59

Female Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.76	-1.00	-0.75	-0.25	0.80	1.51
5dB	-1.76	3.15	3.33	3.67	4.29	4.60
10dB	3.24	7.37	7.48	7.66	7.82	7.78
15dB	8.24	11.11	11.16	11.18	11.05	10.70
20dB	13.24	15.06	15.09	15.06	14.73	14.28

Table 6.47: Quality evaluation of the enhanced speech for various α values for ML estimator based on Laplacian speech model (with VAD): Segmental SNR (Female speech)

Male Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-6.30	-0.25	0.01	0.58	1.93	2.88
5dB	-1.30	4.06	4.29	4.66	5.47	5.76
10dB	3.70	8.01	8.15	8.39	8.75	8.82
15dB	8.70	11.87	11.95	12.01	11.94	11.70
20dB	13.70	15.74	15.77	15.73	15.28	14.73

Male Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-5.94	-0.62	-0.41	0.01	0.99	1.73
5dB	-0.94	3.34	3.48	3.72	4.05	4.14
10dB	4.06	7.23	7.32	7.43	7.49	7.65
15dB	9.06	11.15	11.18	11.16	10.83	10.46
20dB	14.06	15.40	15.41	15.36	14.98	14.43

Male Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0dB	-7.87	-1.98	-1.73	-1.23	0.01	1.68
5dB	-2.87	2.13	2.30	2.62	3.30	4.08
10dB	2.13	5.82	5.94	6.12	6.39	6.49
15dB	7.13	9.79	9.85	9.89	9.75	9.47
20dB	12.13	13.84	13.87	13.84	13.44	12.89

Table 6.48: Quality evaluation of the enhanced speech for various α values for ML estimator based on Laplacian speech model (with VAD): Segmental SNR (Male speech)

Female Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.02	1.58	1.57	1.60	1.78	1.80
5	1.66	1.20	1.20	1.19	1.27	1.33
10	1.29	0.87	0.86	0.86	1.08	1.11
15	0.91	0.63	0.62	0.63	0.81	0.85
20	0.60	0.40	0.40	0.42	0.54	0.60

Female Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.61	1.36	1.36	1.39	1.68	1.72
5	1.31	1.13	1.14	1.17	1.27	1.49
10	1.06	0.90	0.91	0.96	1.12	1.24
15	0.78	0.65	0.66	0.70	0.94	1.00
20	0.56	0.45	0.46	0.49	0.67	0.72

Female Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.12	1.62	1.60	1.56	1.65	1.66
5	1.72	1.25	1.23	1.22	1.33	1.40
10	1.31	0.89	0.88	0.88	1.04	1.08
15	0.96	0.67	0.67	0.67	0.80	0.90
20	0.64	0.46	0.45	0.46	0.54	0.61

Table 6.49: Quality evaluation of the enhanced speech for various α values for ML estimator based on Laplacian speech model (with VAD): Log Likelihood Ratio (Female speech)

Male Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	2.49	1.89	1.87	1.83	1.85	1.63
5	2.01	1.47	1.46	1.45	1.54	1.57
10	1.58	1.14	1.13	1.13	1.27	1.28
15	1.17	0.84	0.83	0.84	0.97	0.99
20	0.81	0.61	0.61	0.62	0.74	0.76

Male Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.38	1.08	1.08	1.09	1.24	1.25
5	1.08	0.82	0.82	0.83	1.01	1.04
10	0.78	0.58	0.58	0.60	0.76	0.80
15	0.53	0.44	0.45	0.48	0.65	0.69
20	0.36	0.29	0.29	0.32	0.41	0.50

Male Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		α=0.2	α=0.3	α=0.5	α=1	α=1.5
0	1.67	1.37	1.37	1.40	1.64	1.74
5	1.35	1.11	1.11	1.13	1.42	1.43
10	1.08	0.80	0.80	0.81	0.99	1.13
15	0.79	0.62	0.61	0.65	0.83	0.85
20	0.55	0.41	0.41	0.42	0.55	0.61

Table 6.50: Quality evaluation of the enhanced speech for various α values for ML estimator based on Laplacian speech model (with VAD): Log Likelihood Ratio (Male speech)

6.5.1 Performance Analysis of the Proposed ML-based Speech Enhancement Algorithm with Laplacian Speech Model

In this section, the equation (6.73) is rewritten to express the enhanced speech $\hat{s} = Gv$, where the noisy speech is multiplied by a gain of G.

$$G = \frac{\hat{s}_{ML-Lap}}{v} = \frac{1 - \lambda_N/2av + \sqrt{(1 - \lambda_N/2av)^2 + 4(1-2\alpha)\lambda_N/v^2}}{2} = \frac{1 - \frac{1}{\sqrt{2\xi\gamma}} + \sqrt{\left(1 - \frac{1}{\sqrt{2\xi\gamma}}\right)^2 + \frac{4(1-2\alpha)}{\gamma}}}{2} \tag{6.88}$$

where $\xi = \frac{\lambda_X}{\lambda_N}$ and $\gamma = \frac{v^2}{\lambda_N}$ are the *a priori* and *a posteriori* SNR respectively.

Next the gain G is depicted as functions of ξ and the instantaneous SNR $\gamma - 1$. Three different figures are shown for three α values of 0.3, 0.5 and 1.

From Figure 6.6, the gain functions converge to 0 dB for high γs with different α values. This means when γ is high, the signal is left unchanged. This leads to the better preservation of the speech comparing to the Gaussian assumption. On the other hand, it keeps the additive noise in the enhanced speech as well.

When the prior SNR is high, α=0.3 has higher gains, while α=0.5 and 1 provide lower gains with lower γ. This means α=0.3 will amplify weak speech with high ξ and α=1 will attenuate them. These amplification with α=0.3 improves the quality of enhanced speech with less distortion and better PESQ scores under the ideal variance estimation (Tables 6.33, 6.34, 6.37 and 6.38).

On the other hand, α=0.5 provides better noise removal than smaller αs (Tables 6.35 and 6.36). This can be explained by the lower gains for the same γ values for α=0.5 than α=0.3. Neither speeches nor noises will be amplified with α=0.5 at any time.

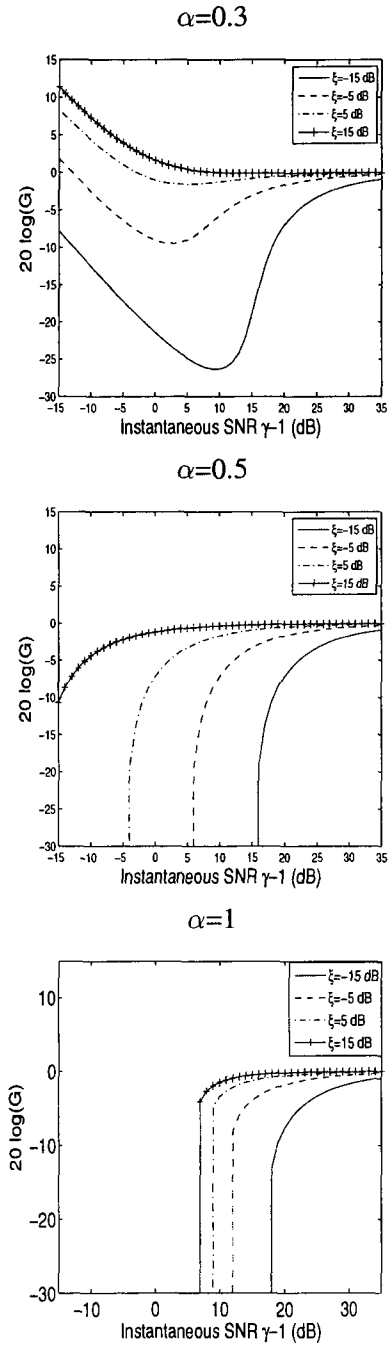

Figure 6.6: The gain function for Maximum Likelihood speech enhancement algorithm with Laplacian speech model

For α=1 or 1.5, more speech/noise is removed from the original noisy signal. Under the ideal prior SNR estimations, it leads to damage of the speech signal while reducing the noise. The quality measures all decrease.

When the noisy signal is used for the speech variance estimation, the prior SNR is estimated with error. Similar to the Gaussian assumption, the estimated prior SNRs tend to be higher than the actual value at the low prior SNR frames. For the high SNR frames, the estimated prior SNRs tend to be lower than the actual value. Also lower prior SNRs lead to larger errors in the prior SNR estimation.

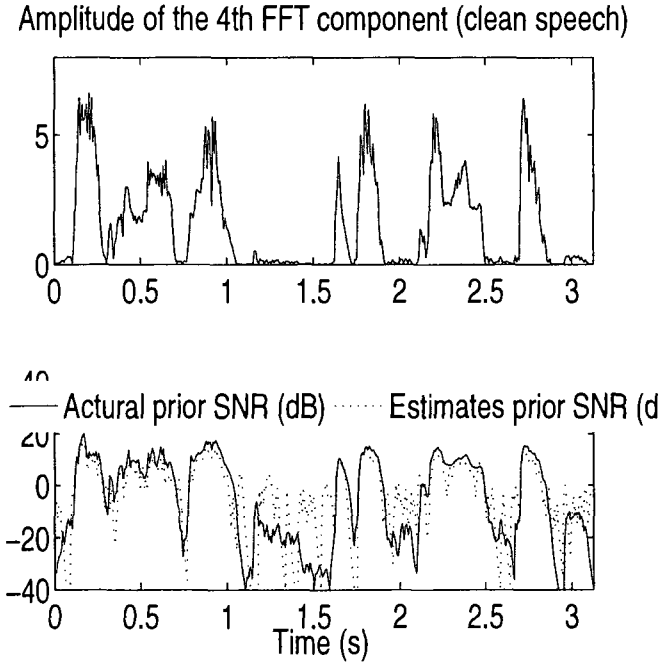

Figure 6.7: A typical comparison between the estimated prior SNR under 0 dB additive noise and the actual prior SNR (Laplacian Assumption)

When α=0.3, these errors in prior SNR estimation can lead to amplification of the errors in the noisy speech. The speech quality will be effected considerably. This effect is shown

by the speech quality scores for α=0.3 (so is α=0.2) in Tables 6.33 to 6.44.

When α=1 is used, the lower estimated SNR at the high SNRs will actually lead to a certain level of attenuation of the speech rather than keep it unchanged. When the instantaneous SNR is low and the prior SNR estimation is less accurate, the algorithm will remove the noise and speech no matter how the prior SNR is estimated. So the results are less effected by the prior SNR estimation at the low SNRs and improved at the high SNRs by slight attenuation to the original signal. The overall speech quality is better than using smaller αs. Tables 6.39 and 6.44 show that α=1 lead to the best overall PESQ scores, although the distortion is still high. The segmental SNRs are also improved comparing to smaller αs, except for the 20 dB case.

With the VAD, the Gaussian assumption leads to better results in all the quality measures than the Laplacian assumption. The noise at low instantaneous SNR frames is mostly removed by VAD in both cases. The Laplacian model keeps the high instantaneous SNR frames unchanged rather than attenuated (the Gaussian model). Thus there is no enhancement at the high SNR frames. The noise residues in these frames become the major drawback for the Laplacian model under such tests.

6.6 Comparison of the Proposed ML-Based Speech Enhancement Algorithms

In this section, we compare the algorithms proposed in this chapter. The ML-based speech enhancement algorithms with Gaussian speech model and Laplacian speech model are compared to the MMSE-based speech enhancement algorithms (with Gaussian speech model). Based on the previous discussions, $\alpha = 0.3$ is used for all algorithms and voice

activity detector is applied.

6.6.1 Performance Analysis

First we compare the PESQ improvements, Segmental SNR and Log likelihood ratio measures. Then we will compare the composite measures.

From Tables 6.51 to 6.56, we can clearly see that the ML-based approach with Gaussian speech model performs comparable with the MMSE-based approach. The PESQ measure is comparable with the MMSE-based approach performs slightly better. The MMSE-based approach provide better noise reduction (segmental SNR) than the ML-based algorithm but also lead to higher distortion (log likelihood ratio).

Between the two proposed ML-based approaches, the approach with Gaussian speech model performs better than the approach with Laplacian speech model. As mentioned before, the main drawback of the Laplacian model-based approach is the non-enhancement for the instantaneous SNR frames regardless of the prior SNR.

From Tables 6.57 to 6.62, a set of composite quality measures are used to evaluate the performance:

$$M_{overall} = 1.594 - 0.512 \cdot LLR + 0.805 \cdot PESQ \tag{6.89}$$

$$M_{back} = 1.634 + 0.063 \cdot segSNR + 0.478 \cdot PESQ \tag{6.90}$$

$$M_{sig} = 3.093 - 1.029 \cdot LLR + 0.603 \cdot PESQ \tag{6.91}$$

where a higher value implies better quality. $M_{overall}$, M_{back} and M_{sig} are the evaluations of the overall quality, the background noise and the speech signal, respectively. The ML-based approach with Gaussian speech model can achieve comparable overall and signal quality with the MMSE-based approach. The background noise level is slightly higher for the ML-based approach.

Female Speech 1

Input SNR	Input PESQ	PESQ score improvement				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.36	+0.39	+0.24	+0.28	+0.22	+0.39
5dB	1.64	+0.45	+0.25	+0.36	+0.26	+0.50
10dB	1.93	+0.49	+0.29	+0.42	+0.31	+0.53
15dB	2.32	+0.47	+0.29	+0.36	+0.27	+0.50
20dB	2.67	+0.45	+0.26	+0.39	+0.24	+0.50

Female Speech 2

Input SNR	Input PESQ	PESQ score improvement				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.60	+0.25	+0.18	+0.19	+0.14	+0.27
5dB	1.87	+0.32	+0.22	+0.24	+0.19	+0.31
10dB	2.17	+0.32	+0.20	+0.24	+0.19	+0.32
15dB	2.44	+0.33	+0.23	+0.32	+0.23	+0.37
20dB	2.73	+0.36	+0.21	+0.30	+0.25	+0.37

Female Speech 3

Input SNR	Input PESQ	PESQ score improvement				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.25	+0.52	+0.30	+0.39	+0.30	+0.61
5dB	1.60	+0.55	+0.31	+0.43	+0.33	+0.63
10dB	1.96	+0.55	+0.32	+0.45	+0.32	+0.60
15dB	2.33	+0.48	+0.34	+0.41	+0.30	+0.52
20dB	2.70	+0.45	+0.27	+0.39	+0.27	+0.49

Table 6.51: Quality evaluation of the enhanced speech for ML and MMSE based speech enhancement algorithms: PESQ score improvement (Female speech)

Male Speech 1

Input SNR	Input PESQ	PESQ score improvement				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.39	+0.29	+0.09	+0.19	+0.17	+0.25
5dB	1.57	+0.41	+0.21	+0.28	+0.22	+0.42
10dB	1.83	+0.47	+0.29	+0.35	+0.30	+0.49
15dB	2.15	+0.42	+0.26	+0.35	+0.23	+0.47
20dB	2.49	+0.34	+0.23	+0.30	+0.24	+0.31

Male Speech 2

Input SNR	Input PESQ	PESQ score improvement				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.63	+0.42	+0.19	+0.39	+0.21	+0.34
5dB	1.95	+0.38	+0.26	+0.39	+0.29	+0.46
10dB	2.29	+0.54	+0.27	+0.55	+0.28	+0.48
15dB	2.68	+0.50	+0.34	+0.46	+0.31	+0.51
20dB	3.11	+0.28	+0.18	+0.28	+0.25	+0.31

Male Speech 3

Input SNR	Input PESQ	PESQ score improvement				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.35	+0.47	+0.29	+0.38	+0.30	+0.51
5dB	1.70	+0.55	+0.35	+0.45	+0.32	+0.56
10dB	2.06	+0.62	+0.37	+0.48	+0.35	+0.66
15dB	2.47	+0.54	+0.37	+0.44	+0.35	+0.53
20dB	2.84	+0.42	+0.29	+0.48	+0.28	+0.48

Table 6.52: Quality evaluation of the enhanced speech for ML and MMSE based speech enhancement algorithms: PESQ score improvement (Male speech)

Female Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	-6.56	1.17	-1.60	-0.08	-2.01	1.92
5dB	-1.56	4.67	2.64	3.84	2.47	5.37
10dB	3.44	8.20	6.73	7.56	6.66	8.58
15dB	8.44	11.72	10.88	11.30	10.80	11.87
20dB	13.44	15.46	15.01	15.27	14.97	15.40

Female Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	-11.78	-3.68	-6.83	-5.12	-7.02	-2.46
5dB	-6.78	0.12	-2.32	-0.99	-2.73	1.14
10dB	-1.78	3.67	1.84	3.03	1.66	4.41
15dB	3.22	7.46	6.02	6.80	5.88	7.82
20dB	8.22	11.16	10.19	10.81	10.15	11.36

Female Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	-6.76	0.41	-2.11	-0.75	-2.35	1.41
5dB	-1.76	4.27	2.13	3.33	2.10	4.91
10dB	3.24	7.87	6.51	7.48	6.34	8.30
15dB	8.24	11.54	10.65	11.16	10.56	11.75
20dB	13.24	15.39	14.82	15.09	14.88	15.44

Table 6.53: Quality evaluation of the enhanced speech for ML and MMSE based speech enhancement algorithms: Segmental SNR (Female speech)

Male Speech 1

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	-6.30	1.35	-1.31	0.01	-1.59	2.19
5dB	-1.30	5.24	2.98	4.29	2.79	5.94
10dB	3.70	8.94	7.29	8.15	7.05	9.34
15dB	8.70	12.48	11.39	11.95	11.19	12.62
20dB	13.70	16.07	15.50	15.77	15.46	15.96

Male Speech 2

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	-5.94	0.46	-1.72	-0.41	-1.87	1.62
5dB	-0.94	4.22	2.49	3.48	2.37	5.00
10dB	4.06	7.89	6.75	7.32	6.65	8.14
15dB	9.06	11.59	10.89	11.18	10.92	11.76
20dB	14.06	15.53	15.21	15.41	15.18	15.52

Male Speech 3

Input SNR	Input SegSNR (dB)	Segmental SNR (dB)				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	-7.87	-0.56	-3.09	-1.73	-3.41	-0.05
5dB	-2.87	3.09	1.08	2.30	0.92	3.71
10dB	2.13	6.60	5.22	5.94	5.05	6.97
15dB	7.13	10.32	9.43	9.85	9.33	10.39
20dB	12.13	14.06	13.53	13.87	13.57	14.11

Table 6.54: Quality evaluation of the enhanced speech for ML and MMSE based speech enhancement algorithms: Segmental SNR (Male speech)

Female Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	2.02	1.45	1.69	1.57	1.72	1.49
5dB	2.66	1.09	1.29	1.20	1.30	1.10
10dB	1.29	0.84	0.94	0.86	0.95	0.79
15dB	0.91	0.55	0.65	0.62	0.67	0.57
20dB	0.60	0.37	0.44	0.40	0.43	0.40

Female Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.61	1.24	1.44	1.36	1.48	1.28
5dB	1.31	1.00	1.13	1.14	1.17	1.03
10dB	1.06	0.83	0.90	0.91	0.90	0.83
15dB	0.78	0.54	0.65	0.66	0.67	0.66
20dB	0.56	0.44	0.48	0.46	0.46	0.45

Female Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	2.12	1.45	1.73	1.60	1.70	1.40
5dB	1.72	1.10	1.36	1.23	1.31	1.07
10dB	1.31	0.79	0.94	0.88	1.02	0.82
15dB	0.96	0.55	0.70	0.67	0.70	0.61
20dB	0.64	0.41	0.45	0.45	0.50	0.44

Table 6.55: Quality evaluation of the enhanced speech for ML and MMSE based speech enhancement algorithms: Log Likelihood Ratio (Female speech)

Male Speech 1

Input SNR	Input LLR	Log Likelihood Ratio				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	2.49	1.72	2.02	1.87	2.12	1.71
5dB	2.01	1.34	1.57	1.46	1.59	1.30
10dB	1.58	1.00	1.20	1.13	1.21	1.07
15dB	1.17	0.77	0.88	0.83	0.91	0.80
20dB	0.81	0.55	0.60	0.61	0.62	0.60

Male Speech 2

Input SNR	Input LLR	Log Likelihood Ratio				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.38	0.97	1.17	1.08	1.17	0.98
5dB	1.08	0.72	0.82	0.82	0.85	0.78
10dB	0.78	0.53	0.62	0.58	0.59	0.61
15dB	0.53	0.32	0.41	0.45	0.43	0.42
20dB	0.36	0.29	0.27	0.29	0.29	0.31

Male Speech 3

Input SNR	Input LLR	Log Likelihood Ratio				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.67	1.12	1.42	1.37	1.40	1.33
5dB	1.35	0.91	1.11	1.11	1.12	1.06
10dB	1.08	0.64	0.84	0.80	0.80	0.84
15dB	0.79	0.50	0.60	0.61	0.61	0.59
20dB	0.55	0.35	0.41	0.41	0.43	0.41

Table 6.56: Quality evaluation of the enhanced speech for ML and MMSE based speech enhancement algorithms: Log Likelihood Ratio (Male speech)

Female Speech 1

Input SNR	Input $M_{overall}$	$M_{overall}$				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.65	2.26	2.02	2.11	1.99	2.24
5dB	2.07	2.72	2.46	2.59	2.46	2.75
10dB	2.51	3.11	2.90	3.05	2.91	3.17
15dB	3.00	3.56	3.36	3.43	3.34	3.57
20dB	3.43	3.92	3.72	3.85	3.72	3.94

Female Speech 2

Input SNR	Input $M_{overall}$	$M_{overall}$				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	2.07	2.45	2.29	2.34	2.23	2.44
5dB	2.41	2.84	2.70	2.71	2.65	2.82
10dB	2.81	3.17	3.04	3.07	3.03	3.17
15dB	3.15	3.55	3.41	3.48	3.40	3.52
20dB	3.50	3.86	3.71	3.80	3.76	3.86

Female Speech 3

Input SNR	Input $M_{overall}$	$M_{overall}$				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.53	2.28	1.96	2.09	1.97	2.37
5dB	2.01	2.76	2.44	2.60	2.47	2.84
10dB	2.50	3.21	2.95	3.08	2.90	3.23
15dB	2.99	3.57	3.38	3.46	3.35	3.58
20dB	3.42	3.92	3.75	3.85	3.72	3.93

Table 6.57: Quality evaluation of the enhanced speech for ML and MMSE based speech enhancement algorithms: $M_{overall}$ (Female speech)

Male Speech 1

Input SNR	Input $M_{overall}$	$M_{overall}$				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.44	2.20	1.90	2.05	1.98	2.19
5dB	1.82	2.62	2.33	2.45	2.36	2.64
10dB	2.27	3.04	2.82	2.90	2.78	3.04
15dB	2.72	3.38	3.18	3.26	3.15	3.39
20dB	3.17	3.66	3.55	3.61	3.54	3.62

Male Speech 2

Input SNR	Input $M_{overall}$	$M_{overall}$				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	2.19	2.75	2.46	2.67	2.46	2.67
5dB	2.62	3.10	2.95	3.06	2.96	3.13
10dB	3.02	3.60	3.34	3.58	3.36	3.51
15dB	3.48	3.99	3.82	3.89	3.78	3.95
20dB	3.92	4.17	4.10	4.17	4.15	4.18

Male Speech 3

Input SNR	Input $M_{overall}$	$M_{overall}$				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.83	2.49	2.19	2.29	2.15	2.41
5dB	2.25	2.94	2.68	2.76	2.64	2.87
10dB	2.71	3.42	3.12	3.23	3.12	3.35
15dB	3.18	3.76	3.57	3.62	3.55	3.71
20dB	3.60	4.04	3.90	4.06	3.88	4.06

Table 6.58: Quality evaluation of the enhanced speech for ML and MMSE based speech enhancement algorithms: $M_{overall}$ (Male speech)

Female Speech 1

Input SNR	Input M_{back} (dB)	M_{back}				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.87	2.54	2.30	2.41	2.26	2.59
5dB	2.32	2.93	2.70	2.83	2.70	2.99
10dB	2.77	3.31	3.12	3.23	3.12	3.35
15dB	3.27	3.71	3.57	3.63	3.55	3.73
20dB	3.76	4.10	3.98	4.06	3.97	4.12

Female Speech 2

Input SNR	Input M_{back} (dB)	M_{back}				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.66	2.29	2.05	2.17	2.02	2.37
5dB	2.10	2.69	2.49	2.58	2.44	2.74
10dB	2.56	3.06	2.88	2.98	2.87	3.10
15dB	3.00	3.43	3.29	3.38	3.28	3.47
20dB	3.46	3.81	3.68	3.76	3.70	3.83

Female Speech 3

Input SNR	Input M_{back} (dB)	M_{back}				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.81	2.51	2.24	2.37	2.22	2.61
5dB	2.29	2.93	2.68	2.81	2.68	3.00
10dB	2.78	3.33	3.13	3.26	3.12	3.38
15dB	3.27	3.70	3.58	3.65	3.56	3.74
20dB	3.76	4.11	3.99	4.06	3.99	4.13

Table 6.59: Quality evaluation of the enhanced speech for ML and MMSE based speech enhancement algorithms: M_{back} (Female speech)

Male Speech 1

Input SNR	Input M_{back} (dB)	M_{back}				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.90	2.52	2.26	2.39	2.25	2.55
5dB	2.30	2.91	2.67	2.79	2.66	2.96
10dB	2.74	3.30	3.11	3.19	3.09	3.33
15dB	3.21	3.65	3.50	3.58	3.47	3.68
20dB	3.69	4.00	3.91	3.96	3.91	3.98

Male Speech 2

Input SNR	Input M_{back} (dB)	M_{back}				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	2.04	2.64	2.40	2.57	2.39	2.68
5dB	2.51	3.01	2.85	2.97	2.85	3.10
10dB	2.98	3.48	3.28	3.45	3.28	3.47
15dB	3.49	3.88	3.76	3.84	3.75	3.90
20dB	4.01	4.23	4.16	4.23	4.20	4.27

Male Speech 3

Input SNR	Input M_{back} (dB)	M_{back}				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.78	2.47	2.22	2.35	2.21	2.52
5dB	2.27	2.90	2.68	2.80	2.66	2.95
10dB	2.75	3.33	3.12	3.22	3.10	3.37
15dB	3.26	3.72	3.59	3.65	3.56	3.72
20dB	3.75	4.08	3.98	4.09	3.98	4.11

Table 6.60: Quality evaluation of the enhanced speech for ML and MMSE based speech enhancement algorithms: M_{back} (Male speech)

Female Speech 1

Input SNR	Input M_{sig}	M_{sig}				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.83	2.66	2.32	2.47	2.28	2.62
5dB	2.38	3.23	2.91	3.06	2.90	3.25
10dB	2.98	3.69	3.46	3.63	3.46	3.76
15dB	3.57	4.21	4.00	4.07	3.97	4.21
20dB	4.07	4.59	4.41	4.53	4.41	4.59

Female Speech 2

Input SNR	Input M_{sig}	M_{sig}				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	2.42	2.93	2.68	2.77	2.62	2.90
5dB	2.83	3.38	3.19	3.19	3.13	3.35
10dB	3.34	3.74	3.60	3.61	3.59	3.74
15dB	3.74	4.21	4.03	4.08	4.01	4.11
20dB	4.15	4.50	4.37	4.45	4.42	4.50

Female Speech 3

Input SNR	Input M_{sig}	M_{sig}				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.70	2.67	2.25	2.44	2.28	2.77
5dB	2.30	3.26	2.85	3.05	2.91	3.34
10dB	2.93	3.79	3.50	3.64	3.41	3.79
15dB	3.54	4.22	3.98	4.06	3.95	4.18
20dB	4.03	4.57	4.42	4.49	4.37	4.56

Table 6.61: Quality evaluation of the enhanced speech for ML and MMSE based speech enhancement algorithms: M_{sig} (Female speech)

Male Speech 1

Input SNR	Input M_{sig}	M_{sig}				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.37	2.61	2.21	2.40	2.28	2.64
5dB	1.95	3.16	2.77	2.94	2.82	3.19
10dB	2.59	3.67	3.40	3.50	3.33	3.65
15dB	3.16	4.08	3.83	3.91	3.81	4.05
20dB	3.74	04.38	4.27	4.31	4.22	4.33

Male Speech 2

Input SNR	Input M_{sig}	M_{sig}				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	2.65	3.33	2.99	3.20	3.00	3.27
5dB	3.17	3.76	3.58	3.66	3.57	3.74
10dB	3.63	4.25	4.00	4.21	4.04	4.14
15dB	4.16	4.68	4.49	4.52	4.45	4.58
20dB	4.61	4.84	4.80	4.84	4.82	4.84

Male Speech 3

Input SNR	Input M_{sig}	M_{sig}				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	2.19	3.04	2.62	2.73	2.65	2.85
5dB	2.69	3.51	3.19	3.25	3.16	3.37
10dB	3.24	4.05	3.69	3.80	3.72	3.87
15dB	3.78	4.39	4.19	4.22	4.17	4.29
20dB	4.24	4.70	4.56	4.67	4.53	4.67

Table 6.62: Quality evaluation of the enhanced speech for ML and MMSE based speech enhancement algorithms: M_{sig} (Male speech)

Tables 6.63 to 6.66 shows the performances of ML and MMSE-based speech enhancement algorithms under non-white noises. F16 noise and the babble noise are used [16].

Preliminary results shows that the proposed ML-based algorithms still perform comparable to the MMSE-based algorithm. The PESQ improvements and Segmental SNR all improve with the proposed ML-based algorithms. The LLR results are improved for all but one case (Babble noise, female case). The LLR results are similar before and after enhancement due to the fact that additive babble noise has frequency characteristics similar to the speech itself. At the same time, the noise reduction is still significant (shown by the segmental SNR measure) and the overall quality is improved (PESQ).

6.6.2 Complexity of Proposed Algorithms

The majority of computational complexity in the MMSE-based algorithms is the calculation of the confluent hypergeometric function (CHF). This function is calculated by the infinite summation as in (6.15). In [9], the infinite summation is calculated with the first 500 terms to achieve desirable accuracy. For r increasing by one, the added term in (6.17) can be calculated as the previous term multiply by $(-\alpha)(-\beta)$ and divided by r two times, which is:

$$\frac{(-\alpha)_r (-\beta)^r}{(r!)^2} = \frac{(-\alpha)_{r-1} (-\beta)^{r-1}}{((r-1)!)^2} \frac{(-\alpha+r-1)(-\beta)}{(r)^2}. \tag{6.92}$$

Thus, we can consider it as 4 multiplications for each term and 2000 multiplications for the calculations of the CHF. With the Gaussian and MMSE approach, only one CHF is needed at each sample. With the Laplacian approach, the infinite summation of the CHF functions are needed [30]. Even though the authors suggested that only first 40 terms of the summation is needed, the computational load is still extremely high.

Input SNR	Input PESQ	PESQ score improvement				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.47	+0.20	+0.09	+0.18	+0.10	+0.22
5dB	1.90	+0.15	+0.10	+0.14	+0.08	+0.16
10dB	2.30	+0.12	+0.09	+0.18	+0.07	+0.19
15dB	2.70	+0.14	+0.07	+0.15	+0.05	+0.15
20dB	3.06	+0.19	+0.09	+0.18	+0.08	+0.18

Input SNR	Input Segmental SNR (dB)	Segmental SNR (dB)				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	-6.58	-0.15	-3.29	-1.80	-3.66	-0.40
5dB	-1.58	2.69	0.88	2.12	0.66	2.75
10dB	3.42	6.05	4.72	5.79	4.45	6.34
15dB	8.42	10.02	9.14	9.65	9.02	10.09
20dB	13.42	14.00	13.74	13.92	13.69	14.16

Input SNR	Input LLR	Log Likelihood Ratio				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	0.95	1.12	1.32	1.17	1.30	1.09
5dB	0.74	0.89	1.01	0.91	1.02	0.87
10dB	0.53	0.54	0.68	0.58	0.70	0.54
15dB	0.36	0.34	0.40	0.36	0.41	0.34
20dB	0.23	0.25	0.31	0.26	0.33	0.23

Table 6.63: Comparison of the objective quality measures for ML and MMSE based speech enhancement algorithms (Female speech 3, babble noise)

Input SNR	Input PESQ	PESQ score improvement				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.66	+0.17	+0.11	+0.18	+0.08	+0.21
5dB	2.02	+0.20	+0.14	+0.15	+0.10	+0.21
10dB	2.37	+0.22	+0.17	+0.18	+0.13	+0.26
15dB	2.74	+0.25	+0.21	+0.24	+0.16	+0.28
20dB	3.13	+0.21	+0.13	+0.17	+0.08	+0.26

Input SNR	Input Segmental SNR (dB)	Segmental SNR (dB)				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	-7.84	-0.96	-3.29	-2.03	-3.80	-0.80
5dB	-2.84	2.49	0.78	2.19	0.59	2.72
10dB	2.16	6.38	5.12	6.10	5.00	6.79
15dB	7.16	9.62	9.03	9.58	9.00	9.79
20dB	12.16	13.66	13.13	13.57	13.07	13.71

Input SNR	Input LLR	Log Likelihood Ratio				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.07	1.02	1.22	1.07	1.20	1.03
5dB	0.82	0.77	0.86	0.81	0.91	0.74
10dB	0.60	0.55	0.59	0.56	0.60	0.53
15dB	0.42	0.41	0.50	0.44	0.51	0.37
20dB	0.28	0.28	0.35	0.27	0.34	0.26

Table 6.64: Comparison of the objective quality measures for ML and MMSE based speech enhancement algorithms (Male speech 3, babble noise)

Input SNR	Input PESQ	PESQ score improvement				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.45	+0.41	+0.26	+0.36	+0.24	+0.41
5dB	1.87	+0.44	+0.32	+0.41	+0.30	+0.46
10dB	2.28	+0.35	+0.27	+0.31	+0.25	+0.36
15dB	2.67	+0.34	+0.27	+0.34	+0.25	+0.35
20dB	3.03	+0.32	+0.25	+0.28	+0.22	+0.32

Input SNR	Input Segmental SNR (dB)	Segmental SNR (dB)				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	-5.94	0.62	-2.06	0.03	-2.31	1.05
5dB	-0.94	4.59	3.01	4.34	2.92	4.77
10dB	4.06	7.20	6.29	6.94	6.15	7.77
15dB	9.06	11.02	10.03	10.95	9.93	11.29
20dB	14.06	15.06	14.49	14.87	14.37	15.17

Input SNR	Input LLR	Log Likelihood Ratio				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.24	1.08	1.12	1.10	1.14	1.03
5dB	0.97	0.83	0.91	0.85	0.92	0.81
10dB	0.70	0.60	0.61	0.00	0.64	0.57
15dB	0.48	0.41	0.46	0.43	0.47	0.40
20dB	0.30	0.27	0.31	0.27	0.33	0.27

Table 6.65: Comparison of the objective quality measures for ML and MMSE based speech enhancement algorithms (Female speech 3, F16 noise)

Input SNR	Input PESQ	PESQ score improvement				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.60	+0.47	+0.34	+0.39	+0.30	+0.48
5dB	1.99	+0.44	+0.35	+0.40	+0.32	+0.46
10dB	2.37	+0.41	+0.32	+0.40	+0.30	+0.44
15dB	2.75	+0.35	+0.27	+0.34	+0.25	+0.37
20dB	3.12	+0.34	+0.26	+0.34	+0.24	+0.38

Input SNR	Input Segmental SNR (dB)	Segmental SNR (dB)				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	-7.10	0.89	-1.29	0.62	-1.45	1.09
5dB	-2.10	3.29	1.18	3.10	0.98	4.21
10dB	2.90	6.80	5.42	6.64	5.29	7.27
15dB	7.90	10.72	10.03	10.55	9.83	10.79
20dB	12.90	14.26	13.83	14.27	13.87	14.41

Input SNR	Input LLR	Log Likelihood Ratio				
		Gaussian, ML		Laplacian, ML		MMSE
		VAD	no VAD	VAD	no VAD	
0dB	1.14	1.02	1.12	1.04	1.10	1.00
5dB	0.89	0.71	0.82	0.73	0.84	0.70
10dB	0.67	0.54	0.59	0.53	0.59	0.52
15dB	0.49	0.37	0.45	0.38	0.46	0.36
20dB	0.33	0.25	0.31	0.26	0.33	0.25

Table 6.66: Comparison of the objective quality measures for ML and MMSE based speech enhancement algorithms (Male speech 3, F16 noise)

On the other hand, the ML-based algorithm is computationally efficient, only 9 multiplications are needed for Gaussian model and 7 for the Laplacian model (at each sample). Also we show that the performance can still be improved with better estimations of the SNRs. A more complex voice activity detector can be applied to further enhance the speech and still lead to less computational complexity than the MMSE approach.

6.7 Conclusion

In this chapter, we first review speech enhancement based on minimizing the MSE in the spectral or loudness domain. Given the complexity of the MMSE approach, we examine the speech enhancement based on the Maximum Likelihood estimation. We proposed a speech enhancement algorithm based on the maximization of the likelihood function of $f\left(|X|^{2\alpha}|Y\right)$ for a given Y and knowledge of the speech and noise variances. The performance of the proposed algorithm is tested for various α values and the quality of the enhanced speech is measured with PESQ, segmental SNR and log-likelihood ratio.

The maximization of $f\left(|X|^{2\alpha}|Y\right)$ is discussed for two statistical modelling assumptions: the Gaussian model and the Laplacian model. First the models are tested with ideal noise and speech variance estimation.

For ideal SNR estimation with both the Gaussian and the Laplacian models, the speech quality is significantly improved for α =0.2, 0.3, 0.5 and 1. The best overall quality and lowest distortion are achieved with α=0.3, which follows the loudness model. For the tested quality measures: PESQ, segmental SNR and log likelihood ratio, the Gaussian model performs comparable with the Laplacian model. The Laplacian model preserves the speech better based on the subjective listening tests while the Gaussian Model removes more background noise. This is due to the fact that Laplacian-based approach leaves the

noisy speech unchanged for high instantaneous SNRs and the Gaussian-based approach attenuate these noisy speech frames.

However, the performance of the proposed algorithm is degraded when we use the actual speech variance estimation. Small α values lead to unnecessary amplification of the noisy signal. Under such conditions, the larger α value of 1 leads to better performance. With larger α, the weak speech frames will be removed instead of amplified. It leads to better noise removal and better overall quality under the noisy environments.

Overall, the proposed speech enhancement algorithm can improve the quality of speech. It proves that the loudness model can provide better performance under the ideal variance and prior SNR estimations. However, the proposed algorithm can perform better with larger α values with the actual variance and prior SNR estimation.

Comparing to the MMSE approaches, the ML approaches achieves consistent improvement over all the SNRs for Gaussian model (without VAD). The ML approaches perform comparable with the MMSE approach with Gaussian assumption and a VAD. The MMSE approach with Laplacian speech model claims further improvement over the MMSE with Gaussian speech model (0.05-0.25 PESQ score overall). On the other hand, the ML approaches have significantly less computational load. The ML approach with Gaussian speech model can provide a efficient substitute for the MMSE approaches.

The ML approach with Laplacian speech model performs slightly worse than the ML with Gaussian approach. The ML approach needs less computations itself, but the Laplacian-based VAD is more complicated. As such the ML algorithm based on Gaussian model is the recommended choice compared to ML Laplacian-based algorithm. The ML approaches are preferable over the MMSE-based algorithms due to the similar performance with considerably less computational complexity.

Chapter 7

Concluding Remarks and Future Considerations

7.1 Conclusions

In this thesis, we have studied the problem of speech enhancement and proposed various algorithms based on the statistical and loudness models of speech.

In chapter 3, MMSE and ML speech enhancement based on the Laplacian model for speech signals has been reviewed. The performance is shown to be limited by the accuracy of the Laplacian parameter estimation under the noisy environment. A recursive version of these algorithms is proposed to estimate the Laplacian parameter using the enhanced speech and then to use these parameters to re-enhance the original noisy speech again. This approach achieves better parameter estimation and hence further improvements of the speech quality.

Considering that the loudness model is better matched to the human hearing system than the spectral signal representation, the fundamental approaches of spectral subtraction

are extended to the loudness domain. As a commonly used representation of loudness is the α power of the speech energy at a certain frequency, we consider loudness subtraction as a speech enhancement approach.

We provide tests for subtraction with different α power values in chapter 4. Simulations show that the quality of enhanced speech can be optimized by using different α values for different input SNRs. This leads us to propose an adaptive-α subtraction model. The simulations show the proposed adaptive-α approach can further improve the performance of spectral subtraction. Furthermore, this adaptive approach showed consistent improvement for all input SNRs.

In chapter 5, we re-examine the fixed α case. We show that the loudness subtraction leads to better results overall compared to the classical spectral subtraction, even though noise residue and unpleasant artifacts are still high in the enhanced signal. Next, we investigate the loudness over-subtraction method. This approach is more aggressive in noise removal. Extensive simulation studies are conducted showing clear improvement over other subtraction type approaches. The loudness over-subtraction model was proven to be more effective at low SNRs.

In chapter 6, Maximum Likelihood (ML) estimation is used to derive a speech enhancement algorithm. The Laplacian and the Gaussian speech models are used separately for comparison. The approaches designed in the loudness domain outperform those in spectral domain under the ideal *a priori* SNR estimations. The Laplacian model leads to better preservation of the speech and the Gaussian model leads to better noise reduction. Also, results fot $\alpha = 0.2$ and 0.3 (commonly used for the loudness model) are better than those for larger αs. This confirms that the loudness model is a valuable tool for designing speech enhancement algorithms. However, this algorithm was shown to be sensitive to the

a priori SNR estimates. When the variances are estimated from the observed signal, α values of 0.5 and 1 performs better. The performance of the ML approach is shown to be comparable to the MMSE approach for Gaussian assumption. The advantage of the proposed ML approach is that the computational load is considerably lower than MMSE-based approaches.

For all the proposed speech enhancement approaches, the recursive MMSE approach presented in chapter 3 provides the best overall improvement after 4 iterations. Its PESQ and segmental SNR results are comparable with loudness over-subtraction (in chapter 5) and Gaussian-based ML approach with VAD (in chapter 6). The recursive MMSE approach has lower distortion than the other two approach. However, the recursive MMSE approach is more complex than the other proposed approach computationally.

The loudness over-subtraction method provides better results than the ML-based approach without VAD. However, the ML-based approach with the VAD provides similar PESQ scores with the loudness over-subtraction approach. The loudness subtraction removes more noise (lower segmental SNR) and introduces more distortion than the ML-based algorithm with VAD. Both the loudness and ML-based algorithms need considerably less computational load than the recursive MMSE approach.

7.2 Suggestions for Future Research

A list of issues arose during the completion of this thesis that merit further consideration:

1. The proposed adaptive-α subtraction, loudness over-subtraction and maximum likelihood approaches all depend on the estimation of the *a prior* SNR. The estimation of this *a prior* SNR is critical for these algorithms and is difficult under the low SNR

environments. Further effort needs to be done to obtain better estimates.

2. The proposed adaptive-α subtraction and loudness over-subtraction shows the need to adjust the speech enhancement algorithm based on the *a prior* SNR. Different α or over-subtraction factor values are needed for different *a prior* SNRs. Our parameter selections were mainly based on the results of the simulations. Even though effective solutions have been proposed for the algorithms, better understanding of the balance between noise removal and distortion is needed for better selection of the parameters.

3. The loudness model used in this thesis is the simplest one based on Steven's power law. The speech enhancement algorithms based on more complex loudness models need to be investigated.

Bibliography

[1] K. R. Castleman, *Digital Image Processing,* Prentice Hall, 1996.

[2] I. S. Gradshteyn and Z. M. Ryzhik, *Tables of Integrals, series and Products, 6th edition,* Academic Press, 2000.

[3] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing,* Prentice Hall, 1989.

[4] A. Papoulis, *Probability, Random Variables, and Stochastic Process, 3rd ed,* WCB/McGraw-Hill, 1991.

[5] H. V. Poor, *An Introduction to Signal Detection and Estimation, 2nd edition,* New York Springer-Verlag, 1994.

[6] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality,* Englewood Cliffs, NJ, Prentice Hall, 1988.

[7] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice,* Prentice-Hall, 2002.

[8] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals,* Prentice-Hall, Englewood Cliffs, NJ, 1978.

[9] S. Zhang and J. Jin, *Computation of Special Functions*, Wiley, 1996.

[10] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models, 2nd ed*, Berlin: Springer Verlag, 1999.

[11] *Single Ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications,* ITU-T Recommendation P.563, May 2004.

[12] *Methods for Subjective Determination of Transmission Quality,* ITU-T Recommendation P.800, August 1996.

[13] *Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs,* ITU-T Recommendation P.830, February 1996.

[14] *Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithm*, ITU-T Recommendation P.835, 2003.

[15] *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs,* ITU-T Recommendation P.862, February 2000.

[16] NOISEX-92 database, http://spib.rice.edu/spib/select_noise.html

[17] Wolfram Mathworld, http://mathworld.wolfram.com/.

[18] J. B. Allen and S. T. Neely, "Modeling the Relation between the Intensity Just-noticeable Difference and Loudness for Pure Tones and Wideband Noise," *Journal of the Acoustical Society of America,* Vol. 102, No. 6, pp. 3628-3646, December 1997.

[19] J.-E. Appell, "Loudness Models for Rehabilitative Audiology", dissertation, Carl von Ossietzky University Oldenburg, 2002.

[20] H. Arslan, A. McCree, and V. Viswanathan, "New Methods for Adaptive Noise Suppression," *International Conference on Acoustics, Speech, and Signal Processing,* Vol. 1, pp. 812-815, 9-12 May 1995.

[21] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech Enhancement based on the Subspace Method," *EEE Transactions on Speech and Audio Processing,* Vol. 8, No. 5, pp. 497-507, September 2000.

[22] R. Badeau, G. Richard, and B. David, "Sliding Window Adaptive SVD Algorithms," *IEEE Transactions on Signal Processing* Vol. 52, No. 1, pp. 1-10, January 2004.

[23] B. Bauer and E. L. Torick, "Researches in Loudness Measurement," *IEEE Transactions on Audio And Electroacoustics*, Vol. AV-14, No. 3, pp. 141-151, September 1966.

[24] A. Bayya and M. Vis, "Objective Measures for Speech Quality Assessment in Wireless Communications," *International Conference on Acoustics, Speech, and Signal Processing,* Vol. 1, pp. 495-498, 7-10 May 1996.

[25] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol. 4, pp. 208-211 April, 1979.

[26] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. 27, No. 2, pp. 113-120, April 1979.

[27] R. Le Bouquin and G. Faucon, "Maximum Likelihood Noise Cancellation with Spectral Constraints," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol. 2, pp. 941-944, April 1991.

[28] S. Buus, H. Müsch, and M. Florentine, "On Loudness at Threshold," *The Journal of the Acoustical Society of America,* Vol. 104, No. 1, pp. 399-410, July 1998.

[29] B. Carnero and A. Drygajlo, "Perceptual Speech Coding and Enhancement Using Frame-Synchronized Fast Wavelet Packet Transform Algorithms," *IEEE Transactions on Signal Processing*, Vol. 47, No. 6, pp. 1622-1635, June 1999.

[30] B. Chen and P. Loizou, "A Laplacian-based MMSE Estimator for Speech Enhancement," *Speech Communication*, Vol. 49, pp. 134-143, February 2007.

[31] S. Chirtmay and M. Tahernezhadi, "Speech Enhancement Using Wiener Filtering," *Acoustics Letters*, Vol. 21, pp. 110-115, 1997.

[32] I. Cohen and B. Berdugo, "Speech Enhancement for Nonstationary Noise Environments," *Signal Processing*, Vol. 81, No. 11, pp. 2403-2418, November 2001.

[33] W. B. Davenport, "An Experimental Study of Speech Wave Probability Distributions," *Journal of Acoust. Soc. America*, Vol. 24, No.4, pp. 390-399, July 1952.

[34] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech Enhancement from Noise: a Regenerative Approach," *Speech Communication*, No. 1, pp. 45-57, February 1991.

[35] S. Dimolitsas, "Objective Speech Distortion Measures and Their Relevance to Speech Quality Assessments," *IEE Proceedings on Communications, Speech and Vision,* Vol. 136, No. 5, pp. 317-324, October 1989.

[36] G. Ding, T, Huang, and B. Xu, "Suppression of Additive Noise Using a Power Spectral Density MMSE Estimator," *IEEE Signal Processing Letters*, Vol. 11, No. 6, pp. 585-588, June 2004.

[37] S. Doclo and M. Moonen, "GSVD-based Optimal Filtering for Single and Multimicrophone Speech Enhancement," *IEEE Transactions on Signal Processing,* Vol. 50, No. 9, pp. 2230-2244, September 2002.

[38] Y. Ephraim and I. Cohen,"Recent Advancements in Speech Enhancement," *The Electrical Engineering Handbook*, CRC press, 2006.

[39] Y. Ephraim, D. Malah, and B.-H. Juang, "On the Application of Hidden Markov Models for Enhancing Noisy Speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 37, No. 12, pp. 1846-1856, December 1989.

[40] Y. Ephraim and H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Trans. Speech and Audio Processing*, Vol. 3, No. 4, pp. 251-266, July 1995.

[41] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum-Mean Square Error Short-time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 32, No. 6, pp. 1109-1121, December 1984.

[42] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 33, No. 2, pp. 443-445, April 1985.

[43] T. H. Falk and W.-Y. Chan, "Non-intrusive GMM-based Speech Quality Measurement," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 125-128, March 18-23, 2005.

[44] N. Fan, "Low Distortion Speech Denoising Using an Adaptive Parametric Wiener Filter," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 309-12, May 17-21, 2004,

[45] M. Feder, A. V. Oppenheim, and E. Weinstein, "Maximum Likelihood Noise Cancellation Using the EM Algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, No. 2, pp. 204-216, February 1989.

[46] S. Gazor and W. Zhang, "A Soft Voice Activity Detector based on a Laplacian-Gaussian Model," *IEEE Transactions on Speech and Audio Processing*, Vol. 11, No. 5, pp. 498-505, September 2003.

[47] S. Gazor and W. Zhang, "Speech Probability Distribution," *IEEE Signal Processing Letters*, Vol. 10, No. 7, pp. 204-207, July 2003.

[48] S. Gazor and W. Zhang, "Speech Enhancement Employing Laplacian-Gaussian Mixture," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 5, pp. 896-904, September 2005.

[49] M. H. Ghoreishi and H. Sheikhzadeh, "A Hybrid Speech Enhancement System Based on HMM and Spectral Subtraction," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol. 3, pp. 1855-1858, June 2000.

[50] Z. Goh, K.-C. Tan, and B. T. Tan "Kalman-filtering Speech Enhancement Method based on a Voice-unvoiced Speech Model," *IEEE Transactions on Speech and Audio Processing,* Vol. 7, No. 5, pp. 510-524, September 1999.

[51] T. Goldstein and A. W. Rix, "Perceptual Speech Quality Assessment in Acoustic and Binaural Applications," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol. 3, pp. 1064-1067, May 2004.

[52] V. K. Goyal, J. Zhang, and M. Vetterli, "Transform Coding with Backward Adaptive Update," *IEEE Trans. Information Theory*, Vol. 46, No. 4, pp. 1623-1633, July 2000.

[53] P. S. K. Hansen, P. C. Hansen, S. D. Hansen and J. A. Sørensen, "On Speech Enhancement Algorithms based on Signal Subspace Methods," *IEEE Nordic Signal Processing Symposium*, pp. 221-224, Aalborg, Denmark, June 1998.

[54] J. H. Hansen and B. L. Pellom, "An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms," *Proceedings of the International Conference on Speech and Language Processing*, Vol. 6, pp. 2819-2822, December 1998.

[55] P. S. K. Hansen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, "Experimental Comparison of Signal Subspace based Noise Reduction Methods," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 101-104, 15-19 March 1999.

[56] P. S. K. Hansen, *Signal Subspace Methods for Speech Enhancement,* Ph.D. thesis (Revised Version), Digital Signal Processing Section, IMM, Technical Univ. of Denmark, 1999.

[57] 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.

[58] R. Hellman and J. Zwislocki, "Some Factors Affecting the Estimation of Loudness," *Journal of the Acoustical Society of America,* Vol. 33, No. 5, pp. 687-694, May 1961.

[59] Y. Hu and P. Loizou, "A Generalized Subspace Approach for Enhancing Speech Corrupted by Colored Noise," *IEEE Transactions on Speech and Audio Processing,* Vol. 11, No. 4, pp. 334-341, July 2003.

[60] Y. Hu and P. Loizou, "A Perceptually Motivated Subspace Approach for Speech Enhancement," *Proc. Inter. Conf. on Spoken Language Processing (ICSLP)*, pp. 1797-1800, 2002.

[61] Y. Hu and P. Loizou, "Evaluation of Objective Measures for Speech Enhancement," *Proc. INTERSPEECH*, September 2006.

[62] J. Huang and Y. Zhao, "A DCT-Based Fast Signal Subspace Technique for Robust Speech Recognition," *IEEE Trans. Speech and Audio Processing*, Vol. 8, No. 6, pp. 747-751, November 2000.

[63] L. E. Humes and W. Jesteadt, "Models of the Effects of Threshold on Loudness Growth and Summation," *Journal of the Acoustical Society of America*, Vol. 90, No. 4, pp. 1933-1943, October 1991.

[64] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 23, No. 1, pp. 67-72, February 1975.

[65] F. Jabloun, and B. Champagne, "Incorporating the Human Hearing Properties in the Signal Subspace Approach for Speech Enhancement," *IEEE Transactions on Speech and Audio Processing*, Vol. 11, No. 6, pp. 700-708, November 2003.

[66] S. H. Jensen, P. C. Hansen, S. D. Hansen and J. A. Sorensen, "Reduction of Broadband Noise in Speech by Truncated QSVD," *IEEE Transactions on Speech and Audio Processing,* Vol. 3, No. 6,pp. 439-448, November 1995.

[67] S. H. Jensen, J. P. Kargo, C. A. Rodbro, and K. V. Sorensen, "Subspace-based Speech Enhancement with Rank-deficient Prewhitening," *IEEE Workshop on Speech Coding*, pp. 166-168, 6-9 October 2002.

[68] M. T. Johnson, A. C. Lindgren, R. J. Povinelli, and X. Yuan, "Performance of Nonlinear Speech Enhancement Using Phase Space Reconstruction," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol. 1, pp. 920-923, 6-10 April 2003.

[69] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE J. Selected Areas in Communications*, Vol. 6, No. 2, pp. 314-323, February 1988.

[70] B. H. Juang, "The Past, Present, and Future of Speech Processing," *IEEE Signal Processing Magazine*, vol. 15, pp. 24-48, May 1998.

[71] G. S. Kang and L. J. Fransen, "Quality Improvement of LPC-processed Noisy Speech by Using Spectral Subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 37 No. 6, pp. 939-942, June 1989.

[72] M. Karjalainen, "A new Auditory Model for the Evaluation of Sound Quality of Audio Systems," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 10, pp. 608-611, April 1985.

[73] J. U. Kim, S. G. Kim, and C. D. Yoo, "The Incorporation of Masking Threshold to Subspace Speech Enhancement," *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 76-79, April 2003.

[74] W. Kim, S. Kang, and H. Ko, "Spectral Subtraction based on Phonetic Dependency and Masking Effects," *IEE Proceedings on Vision, Image and Signal Processing* , Vol. 147, No. 5, pp. 423-427, October 2000.

[75] D.-S. Kim and A. Tarraf, "Perceptual Model for Non-intrusive Speech Quality Assessment," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol. 3, pp. 1060-1063, May 2004.

[76] D. Klatt, "Prediction of Perceived Phonetic Distance from Critical-band Spectra: a First Step," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1278-1281, May 1982.

[77] H. Kobatake, K.Gyoutoku, and S. Li, "Enhancement of Noisy Speech by Maximum Likelihood Estimation," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol. 2, pp. 973-976, April 1991.

[78] K. H. Lam, O. C. Au, C. C. Chan, K. F. Hui, and S. F. Lau, "Objective Speech Quality Measure for Cellular Phone," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 487-490, May 1996.

[79] H. Lev-Ari and Y. Ephraim, "Extension of the Signal Subspace Speech Enhancement Approach to Colored Noise," *IEEE Signal Processing Letters*, Vol. 10, No. 4, pp. 104-106, April 2003.

[80] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, Vol. 67, No. 12, pp. 1586-1604, December 1979.

[81] J. S. Lim "Evaluation of a Correlation Subtraction Method for Enhancing Speech Degraded by Additive White Noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 471-472, October 1978.

[82] J. P. A. Lochner and J. F. Burger, "Form of the Loudness Function in the Presence of Masking Noise," *Journal of the Acoustical Society of America*, Vol. 33, No. 12, pp. 1705-1707, December 1961.

[83] P. Lockwood and J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and projection, for robust recognition in cars", *Speech Communications*, Vol. 11, No. 2-3, pp. 215-228, June 1992.

[84] R. Martin, "Noise Power Spectral Density Estimation based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 5, pp. 504-512, July 2001.

[85] R. Martin, "Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 5, Part 2, pp. 845-856 September 2005.

[86] R. J. McAulay and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Transactions on Acoustics, Speech and Signal Processing,* Vol. 28, No. 2, pp. 137-145, April 1980.

[87] S. M. McOlash, R. J. Niederjohn, and J. A. Heinen, "A Spectral Subtraction Method for the Enhancement of Speech Corrupted by Nonwhite, Nonstationary Noise," *Proceedings of the 1995 IEEE IECON 21st International Conference on Industrial Electronics, Control, and Instrumentation,* Vol. 2, pp. 872-877, November 1995.

[88] B. C. J. Moore and B. R. Glasberg, "Formulae Describing Frequency Selectivity as a Function of Frequency and Level, and Their Use in Calculating Excitation Patterns," *Hearing Research,* Vol. 28, pp. 209-225, 1987.

[89] N. Morgan and H. A. Bourlard, "Neural Networks for Statistical Recognition of Continuous Speech," *Proceedings of the IEEE* Vol. 83, No. 5, pp. 742-772, May 1995.

[90] S. Nandkumar and J. H. Hansen, "Speech Enhancement based on a New Set of Auditory Constrained Parameters," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol. 1, pp. 1-4, April 1994.

[91] S. Nandkumar and J. H. Hansen, "Dual-channel Iterative Speech Enhancement with Constraints on an Auditory-based Spectrum," *IEEE Transactions on Speech and Audio Processing,* Vol. 3, No. 1, pp. 22-34, January 1995.

[92] R. J. Niederjohn, T. J. Svoren, and J. A. Heinen, "Intelligibility Enhancement of

Noise-Corrupted Speech Based on Formant Tracking Involving Prefiltering ," *Proceedings of the International Conference on Industrial Electronics, Control, Instrumentation, and Automation, Power Electronics and Motion Control,* Vol. 3, pp. 1336-1341, November 1992.

[93] T. Petersen and S. Boll, "Acoustic Noise Suppression in the Context of a Perceptual Model," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 6, pp. 1086-1088, April 1981.

[94] A. Rezayee and S. Gazor, "An Adaptive KLT Approach for Speech Enhancement," *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 2, pp. 87-95, February 2001.

[95] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)-a New Method for Speech Quality Assessment of Telephone Networks and Codecs," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 749-752, May 2001.

[96] A. W. Rix and M. P. Hollier, "The Perceptual Analysis Measurement System for Robust End-to-End Speech Quality Assessment," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 1515-1518, June 2000.

[97] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective Assessment of Speech and Audio Quality - Technology and Applications," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 14, No. 6, pp. 1890-1901, November 2006.

[98] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective Perceptual Quality Measures for the Evaluation of Noise Reduction Schemes," *9th International Workshop on Acoustic Echo and Noise Control,* Eindhoven, pp. 169-172, September 2005.

[99] B. L. Sim, Y. C. Tong, J. S. Chan, and T. T. Chin, "A Parametric Formulation of the Generalized Spectral Subtraction Method," *IEEE Transactions on Speech and Audio Processing,* Vol. 6, No. 4, pp. 328-337, July 1998.

[100] J. O. Smith and J. S. Abel, "Bark and ERB Bilinear Transforms," *IEEE Transactions on Speech and Audio Processing,* Vol. 7, No. 6, pp. 697-708, November 1999.

[101] J. O. Smith and J. S. Abel, "The Bark Bilinear Transform", *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, October 1995. Available online at http://ccrma.stanford.edu/ jos/gz/bbtmh.tgz.

[102] I. Y. Soon and S. N. Koh, "Low Distortion Speech Enhancement," *IEE Proceedings-Vision, Image and Speech Processing*, Vol. 147, No. 3, pp. 247-253, June 2000.

[103] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy Speech Enhancement Using Discrete Cosine Transform," *Speech Communication*, Vol. 24, pp. 249-257, June 1998.

[104] S. S. Stevens, "On the Psychophysical Law," *Psychological Reviews*, Vol. 64, pp. 153-181, May 1957.

[105] S. S. Stevens, "Power-Group Transformations under Glare, Masking, and Recruitment," *Journal of the Acoustical Society of America*, Vol. 39, No. 4, pp. 725-735, April 1966.

[106] F. Toledo, P. Loizou, and A. Lobo, "Subspace and Envelope Subtraction Algorithms for Noise Reduction in Cochlear Implants," *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vol. 3, pp. 2002-2005, September 2003.

[107] R. M. Udrea and S. Ciochina, "Speech Enhancement Using Spectral Oversubtraction and Residual Noise Reduction," *International Symposium on Signals, Circuits and Systems,* Vol. 1, pp. 165-168, July 2003.

[108] C. Uhl and M. Lieb, "Experiments with an Extended Adaptive SVD Enhancement Scheme for Speech Recognition in Noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 281-284, May 2001.

[109] N. Virag, "Single Channel Speech Enhancement based on Masking Properties of the Human Auditory System," *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 2, pp. 126-137, March 1999.

[110] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE Journal on Selected Areas in Communications,* Vol.10, No. 5, pp. 819-829, June 1992.

[111] F. Xie and D. V. Compernolle, "Speech Enhancement by Spectral Magnitude Estimation — A Unifying Approach," *Speech Communication*, Vol. 19, No. 2, pp. 89-104, August 1996.

[112] B. Yang, "Projection Approximation Subspace Tracking," *IEEE Trans. Signal Processing*, Vol. 43, No. 1, pp. 95-107, January 1995.

[113] M. Yeary and P. Loizou, "Adaptive Filtering for Speech Enhancement," *Ninth DSP Workshop*, Hunt, TX, October 2000.

[114] C. H. You, S. N. Koh and S. Rahardja, "β-order MMSE Spectral Amplitude Estimation for Speech Enhancement," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 4, pp. 475-486, July 2005.

[115] R. Zelinski and P. Noll, "Adaptive Transform Coding of Speech Signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 25, No. 4, pp. 299-309, August 1977.

[116] W. Zha and W.-Y. Chan, "A Data Mining Approach to Objective Speech Quality Measurement," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 461-464, May 2004.

[117] W. Zhang and S. Gazor, "Statistical Modelling of Speech Signals," *6th International Conference on Signal Processing*, Vol. 1, pp. 480-483, August 2002.

[118] W. Zhang, "Speech Statistical Modelling and its Application in Voice Activity Detector and Speech Enhancement," M.Sc.(Eng.) Thesis, Queen's University, 2002.

[119] W. Zhang and Tyseer Aboulnasr, "A Recursive Speech Enhancement System Employing Laplacian Speech model," *IEEE International Workshop on Haptic Audio Visual Environments and their Applications*, Ottawa, Canada, October 2004.

[120] W. Zhang and Tyseer Aboulnasr, "Speech Enhancement Employing Loudness Subtraction and Over-Subtraction," *Canadian Acoustical Association Annual Conference*, Montreal, Canada, October 2007.

Appendix : Speech and Noise Files Specification

The speech files are selected from the TIMIT database. The sampling rate is 8 KHz. The speech files used are listed below with its file number in database and the text of the speech files.

Female 1 (si1386): In wage negotiation, the industry bargains as a unit with a single union.

Female 2 (si2016): Heave on those ropes the boat's come unstuck.

Female 3 (sx36): Only the most accomplished artist obtain popularity.

Male 1 (sa1): She had your dark suit in greasy wash water all year.

Male 2 (sa2): Don't ask me to carry an oily rag like that.

Male 3 (si1039): He has never, himself, done anything for which to be hated – which of us has?

The noise samples used are white noise (generated by the Matlab), babble noise and F16 cockpit noise (from the NOISEX-92 database, NOISE-ROM-0 signal 011 and 020, also available at http://spib.rice.edu/spib/select_noise.html). The NOISEX-92 database was originally sampled at 19.98 KHz. It is re-sampled with Matlab to 8 KHz.

Babble noise (NOISE-ROM-0 signal.011): The source of this babble is 100 people

speaking in a canteen. The room radius is over two meters; therefore, individual voices are slightly audible.

F16 cockpit noise (NOISE-ROM-0 signal.020): The noise was recorded at the co-pilot's seat in a two-seat F-16, traveling at a speed of 500 knots, and an altitude of 300-600 feet. It was found that the flight condition had only a minor effect on the noise. The reproduced noise can therefore be considered to be representative.