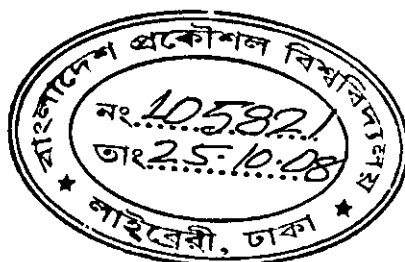


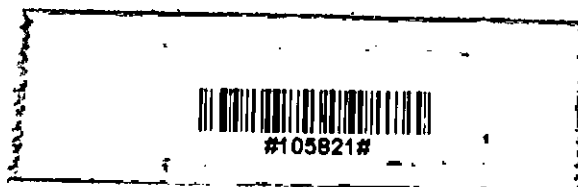
A Hybrid Speech Enhancement Method using Optimal Dual Gain Filters and EMD Based Post Processing

by

Taufiq Hasan Al Banna



MASTER OF SCIENCE IN ELECTRICAL AND ELECTRONIC
ENGINEERING




Department of Electrical and Electronic Engineering

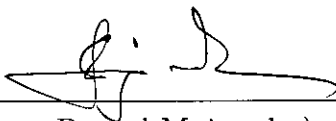
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY


May 2008


The thesis entitled “A Hybrid Speech Enhancement Method using Optimal Dual Gain Filters and EMD Based Post Processing” submitted by Taufiq Hasan Al Banna Roll No.: 100606226P, Session: October, 2006 has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Master of Science in Electrical and Electronic Engineering on May 31, 2008.

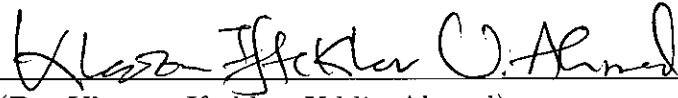
BOARD OF EXAMINERS

1. 

(Dr. Md. Kamrul Hasan)
Professor
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka - 1000. **Chairman**
2. 

(Dr. Satya Prasad Majumder)
Professor and Head
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka - 1000. **Member**
(Ex-officio)
3. 

(Dr. Newaz Muhammad Syfur Rahim)
Associate Professor
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka - 1000. **Member**
4. 

(Dr. Mohammed Imamul Hassan Bhuiyan)
Assistant Professor
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka - 1000. **Member**
5. 

(Dr. Khawza Iftekhhar Uddin Ahmed)
Assistant Professor and Head
Department of Electrical and Electronic Engineering
United International University, Dhaka - 1209. **Member**
(External)

Declaration

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

Signature of the candidate



(Taufiq Hasan Al Banna)

Dedication

To my beloved parents.

Acknowledgements

I would like to commence this thesis by expressing my gratitude towards Almighty *Allah*, *Who* blessed me with *His* endless mercy, *Filled* me with diligence and perseverance, and most importantly, *Created* kindness in so many people who provided me with the necessary support I required. Among those people, the one person who stands out above the rest is my supervisor, Professor Md. Kamrul Hasan. I am greatly indebted to him for his guidance, limitless patience and constant support that was critical for completing this research. It was an honor working with him. I am also greatly indebted to Professor Md. Ali Choudhury for his constant support, affection and encouragement, which was necessary for continuing this work. I am especially thankful to Ms. Lutfa Akter, for writing such a comprehensive M.Sc. thesis of her own that was extremely helpful to me at every stage of the work. I would also like to thank Dr. Asif Mohammad Zaman, for providing me with the latest research papers online. This was a crucial support that was necessary to keep pace with the state-of-the art technology.

And last but not the least, I would be eternally grateful to my parents because of their incessant passion, encouragement and support. I humbly dedicate this thesis to them.

Contents

Acknowledgements	iv
List of abbreviations	viii
List of symbols	ix
List of Figures	x
List of Tables	xiv
Abstract	xv
1 Introduction	1
1.1 Speech enhancement	1
1.2 Objective of this research	3
1.3 Organization of this thesis	4
2 Stochastic model based speech enhancement: A review	5
2.1 Introduction	5
2.2 Short-time spectral subtraction	6
2.3 Wiener filtering and its variants	8
2.3.1 Wiener filter in DCT domain	9
2.3.2 The dual gain Wiener filter	11
2.4 Minimum mean square error (MMSE) estimators	13
2.4.1 MMSE estimator in the DFT domain	15
2.4.2 MMSE estimator in the DCT domain	15
2.4.3 Conclusion	16

3	The proposed dual MMSE estimator	17
3.1	The dual MMSE estimator	18
3.2	Properties of the gain $G_{\text{MMSE}+}$	24
3.3	Properties of the gain $G_{\text{MMSE}-}$	26
3.4	Performance comparison and discussion	28
3.4.1	Experiment using generated Gaussian sequences	28
3.4.2	Experiments on speech files	29
3.5	Conclusion	31
4	Post processing using the empirical mode decomposition	35
4.1	Basics of EMD	36
4.1.1	Intrinsic mode functions	36
4.1.2	The sifting algorithm	36
4.2	Separation of Musical Noise using EMD	38
4.3	Proposed method	40
4.3.1	Conservation of energy in EMD	40
4.3.2	Statistical model of IMF local variance	41
4.3.3	Optimum gain	42
4.3.4	Considering speech presence uncertainty	43
4.3.5	Noise variance estimation	44
4.4	Simulation results	44
4.4.1	Experimental details	44
4.4.2	Performance Evaluation and Discussion	46
4.5	Conclusion	50
5	Performance analysis of the hybrid method	55
5.1	Introduction	55
5.2	Experimental details	55
5.3	Performance comparison and discussion	56
5.4	Conclusion	57
6	Conclusion	61
6.1	Summary	61
6.2	Future works	62

A Important derivations	68
A.1 Derivation of the dual gain Wiener	68
A.2 MMSE estimator in the DCT domain	73
A.3 Derivation of the dual MMSE estimators	75
A.3.1 The marginal densities $p(Y_k, H_+)$ and $p(Y_k, H_-)$	75
A.3.2 The dual MMSE estimators	78
A.4 The Ephraim and Malah suppression rule	80
 Bibliography	 64

List of abbreviations

AvgSegSNR	Average Segmental signal to noise ratio
COMP	Composite speech quality measure
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
DGW	Dual gain Wiener
DMMSE	Dual minimum mean square error
EMD	Empirical mode decomposition
EMSR	Ephraim-Malah suppression rule
IMF	Intrinsic mode function
MAP	Maximum a posteriori
MMSE	Minimum mean square error
MSE	Mean square error
PESQ	Perceptual evaluation of speech quality
SNR	Signal to noise ratio

List of Symbols

$x(n)$	The n th time sample of the clean signal
$y(n)$	The n th time sample of the noisy signal
$d(n)$	The n th time sample of the noise
$\mathbf{x}[n]$	A vector containing the clean signal time samples
$\mathbf{y}[n]$	A vector containing the noisy signal time samples
$\mathbf{d}[n]$	A vector containing the noise time samples
X_k	Clean signal DCT/DFT coefficient in the k th bin
Y_k	Noisy signal DCT/DFT coefficient in the k th bin
D_k	Noise DCT/DFT coefficient in the k th bin
$x_i(n)$	The i th IMF samples of the clean signal
$y_i(n)$	The i th IMF samples of the noisy signal
ξ_k	A <i>priori</i> SNR in the k th frequency index
γ_k	A <i>posteriori</i> SNR in the k th frequency index
G_W	The Wiener filter gain
$E\{\cdot\}$	The expectation operator
$\text{sgn}(\cdot)$	The signum function
$\text{erf}(\cdot)$	The error function
$\text{erfc}(\cdot)$	The complementary error function
$U(\cdot)$	The unit step function
$\mathcal{N}(\mu, \sigma)$	A normally distributed random variable with mean μ and standard deviation σ
p_k	Constructive or destructive interference indicator for the k th frequency index
\hat{X}^W	The Wiener estimator
\hat{X}^{DGW}	The dual gain Wiener estimator
\hat{X}^{DMMSE}	The dual MMSE estimator
σ_x^2	Clean signal variance in the k th frequency index
σ_d^2	Noise variance in the k th frequency index
σ_y^2	Noisy signal variance in the k th frequency index
$\lambda_x(i, k)$	Clean signal variance in the i th IMF and k th frame.
$\lambda_y(i, k)$	Noisy signal variance in the i th IMF and k th frame.
$\lambda_d(i, k)$	Noise variance in the i th IMF and k th frame.

List of Figures

2.1	Plots of the the dual gain Wiener (DGW) filters G_{W+} and G_{W-} proposed by [3] as given in (2.18) and (2.19). The gains are plotted against the variation of the <i>a priori</i> SNR values. The conventional Wiener gain G_W is also shown in the same plot.	12
2.2	The parametric gain curves of the Empraim-Malah suppression rule (EMSR).	16
3.1	(a) Valid regions of X_k and Y_k for the events H_+ and H_- , (b) The joint density function $p_{XY}(x_k, y_k)$, (c) The joint density function $p(x_k, y_k, H_+)$ and (d) The joint density function $p(x_k, y_k, H_-)$. . .	20
3.2	(a) The probability density function $p(Y_k, H_+)$ and (b) the probability density function $p(Y_k, H_-)$ when $\sigma_d = \sigma_x = 1$	22
3.3	Parametric gain curves describing (a) MMSE gain function in the constructive case, G_{MMSE+} (solid lines), (b) the gain function due to Emphraim-Malah (EMSR) (dash starred line), (c) the gain G_{W+} from the dual gain Wiener (dashed lines) as in (2.18) and (d) the Wiener gain function.	24
3.4	Parametric gain curves plotted against ξ_k describing (a) MMSE gain function in the constructive case, G_{MMSE+} (solid lines), (b) the gain G_{W+} from the dual gain Wiener (dashed lines) as in (2.18), and (c) the Wiener gain function (dotted lines).	25
3.5	Parametric gain curves describing (a) MMSE gain function in the destructive case, G_{MMSE-} (solid lines), (b) the gain G_{W-} from the dual gain Wiener as in (2.19), and (c) the Wiener gain function (dotted lines).	26

3.6	Parametric gain curves plotted against ξ_k describing (a) MMSE gain function in the destructive case, $G_{\text{MMSE-}}$ (solid lines), (b) the gain G_{W-} from the dual gain Wiener as in (2.18), and (c) the Wiener gain function (dotted lines).	27
3.7	Theoretical performance comparison of the conventional Wiener, dual gain Wiener and the proposed dual gains in the known polarity case.	28
3.8	Performance comparison of conventional Wiener (dashed lines), dual gain Wiener (DGW) (dotted lines) and the Proposed dual MMSE (DMMSE) estimator (solid lines) with respect to improvement in overall SNR, average segmental SNR, composite speech quality measure (COMP) and PESQ scores. Polarity estimator accuracy was assumed to be 100% and the averaging parameter α was set to 0.8.	30
3.9	Performance comparison of Wiener (—), DGW (....) and DMMSE estimator (—) for $\alpha = 0.9$. $A_p = 100\%$ for Figs. (a)-(d) and $A_p = 80\%$ for Figs. (e)-(h)	33
3.10	Performance comparison of Wiener (—), DGW (....) and DMMSE estimator (—) for $A_p = 100\%$. $\alpha = 0.98$ for Figs. (a)-(d) and α is variable for Figs. (e)-(h)	34
4.1	(a) A sequence of musical noise. (b) The energy distribution of musical noise in different IMFs (ratio of the i th IMF variance to the overall signal variance).	39
4.2	Scatter plot of $\lambda_y(i, k)$ vs. $\lambda_x(i, k) + \lambda_d(i, k)$ in Log scale.	41
4.3	Average objective quality measures with different input SNRs; (a), (b): white noise; (c), (d): babble noise.	47
4.4	Average objective quality measures with different input SNRs. Results obtained from ‘white’ and ‘babble’ noise are averaged.	48
4.5	Average objective quality measures with different input SNRs. Results obtained from ‘white’ and ‘babble’ noise are averaged.	49



4.6	Enhancement results for the male utterance “Heels place emphasis on the long legged silhouette”. Time domain plots of (a) clean speech, (b) noisy speech (10dB), (c) enhanced using $W_n(\alpha)$ and (d) enhanced using P- $W_n(\alpha)$	51
4.7	Enhancement results for the male utterance “Heels place emphasis on the long legged silhouette”. Spectrogram plots of (a) clean speech, (b) noisy speech (10dB), (c) enhanced using $W_n(\alpha)$ and (d) enhanced using P- $W_n(\alpha)$	52
4.8	Enhancement results for the male utterance “Heels place emphasis on the long legged silhouette”. Time domain plots of (a) clean speech, (b) noisy speech (10dB), (c) enhanced using $MAP(\alpha)$ and (d) enhanced using P- $MAP(\alpha)$	53
4.9	Enhancement results for the male utterance “Heels place emphasis on the long legged silhouette”. Spectrogram plots of (a) clean speech, (b) noisy speech (10dB), (c) enhanced using $MAP(\alpha)$ and (d) enhanced using P- $MAP(\alpha)$	54
5.1	Performance comparison of the DMMSE, DGW, Wiener and EM Log STSA with respect to improvement in average Segmental SNR for an input SNR range of 0 dB to 25 dB. The proposed post filtering method is applied to the DMMSE, Wiener and EM Log STSA methods indicated by the +EMD notation. DMMSE+EMD indicates the proposed hybrid method. A polarity estimator of accuracy 100% and 80% was used in (a) and (b), respectively.	58
5.2	Performance comparison of the DMMSE, DGW, Wiener and EM Log STSA with respect to improvement in Overall SNR for an input SNR range of 0 dB to 25 dB. The proposed post filtering method is applied to the DMMSE, Wiener and EM Log STSA methods indicated by the +EMD notation. DMMSE+EMD indicates the proposed hybrid method. A polarity estimator of accuracy 100% and 80% was used in (a) and (b), respectively.	59



5.3	Performance comparison of the DMMSE, DGW, Wiener and EM Log STSA with respect to improvement in PESQ scores for an input SNR range of 0 dB to 25 dB. The proposed post filtering method is applied to the DMMSE, Wiener and EM Log STSA methods indicated by the +EMD notation. DMMSE+EMD indicates the proposed hybrid method. A polarity estimator of accuracy 100% and 80% was used in (a) and (b), respectively.	60
A.1	Vector Representation of $\mathbf{Y}_k = \mathbf{X}_k + \mathbf{D}_k$	81

List of Tables

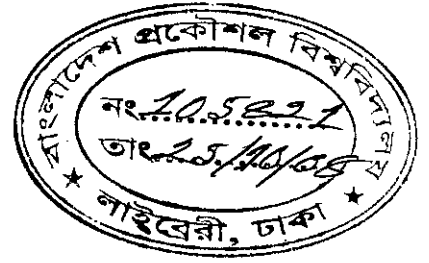
4.1	The comparison category rating (CCR) scale	46
4.2	The mean opinion score (MOS) scale	46
4.3	Results from the Listening test in the CCR scale. A positive value indicates preference for the proposed method	50
4.4	The Mean Opinion Score Experimental Results for white and babble noise of 5 and 10dB SNR	50

Abstract

Several speech enhancement methods have been developed in the past few decades, each method having its advantages and drawbacks. Despite their effectiveness in noise reduction, most conventional algorithms introduce noticeable distortion and annoying musical noise artifact in the enhanced speech. The purpose of this work is to develop a hybrid method of speech enhancement to attain an improved noise reduction performance accompanied by a reduction of the musical noise in the enhanced speech in two different stages.

Conventional spectral subtraction based algorithms assume that noise transform coefficients are always additive with the signal, whereas it may also be subtractive. Therefore, to deal with the first problem of noise reduction, a minimum mean square error (MMSE) estimator is derived considering both the additive and subtractive effect of noise in the discrete cosine transform (DCT) domain. The performance of the new estimator is compared to a previously reported work, that also considers these two cases in the DCT domain, and superior results in terms of signal to noise ratio (SNR) and mean squared error (MSE) are obtained. Since the approach provides two different gains for the two cases, they are termed as dual gain filters.

Dealing with the problem of residual noise suppression, a new post-filtering technique is proposed utilizing the empirical mode decomposition (EMD). An optimum gain function in MMSE sense, is derived for short intrinsic mode function (IMF) segments. Finally, the proposed dual gain filters are used to enhance various noisy speech utterances and the new post filtering algorithm is applied for residual noise suppression. The performance of the proposed two stage hybrid technique is compared to well known speech enhancement algorithms and superior results in both objective and subjective quality indices are obtained.



Chapter 1

Introduction

1.1 Speech enhancement

In audio recording, one can never avoid the menacing and ubiquitous sound known as 'noise'. Simply put, it is the sound that inadvertently gets entangled with our desired 'signal', which can be speech, music or any other audio of interest. Only in a sound-proof (low noise) recording studio, one may capture an audio that can be considered 'clean' for all practical purposes. Regrettably, when we require to capture a sound in real-life circumstances, a recording studio is far from being readily available. While waiting for a bus in a busy afternoon, or inside a cacophonous cafeteria, expecting to find a low noise environment before making a phone call is rather visionary. Thus, we must accept the reality that if we want to make recordings of speech, for the purpose of storage, recognition or transmission through a mobile phone network, it would inescapably become 'noisy'. Speech enhancement, is thus essential.

The term speech enhancement, might be rather confounding, if taken literally. Since 'enhancement' deals with improvement of quality, one might wonder what exactly is understood by the quality of a speech¹. In the speech enhancement problem, two principal criteria are used to measure the goodness of speech signals, namely, quality and intelligibility. While the quality of a speech signal deals with its clarity, nature of distortion and amount of background noise, intelligibility deals with the percentage of words that can be clearly understood. It may sound unlikely, but a better quality of speech does not always guarantee a

¹Obviously, we are not referring to the improvement of the speaker's accent, tone or language usage

higher intelligibility; these criteria are independent of each other. Most speech enhancement systems improve the quality of the speech signal at the expense of a reduction of intelligibility. Listeners can usually extract more information from the noisy signal than from the enhanced signal if they listen to it carefully. However, listening to the noisy signal for a long time causes discomfort², which can be reduced by the enhancement algorithms improving only the quality of speech. The intelligibility of an enhanced speech are often assessed using automatic speech recognition tests, since listening sessions with live subjects are expensive and time consuming. Unfortunately, both quality and intelligibility are difficult to quantify and express in a closed form that is amenable to mathematical optimization. Thus, the design of speech enhancement systems is often based on mathematical measures that are somehow believed to be correlated with the quality and/or intelligibility of the speech signal.

“If you enhance a noisy speech, it sounds even worse” - remarked one of my colleagues; quoting the line from a rather frustrated speech enhancement researcher. Evidently, this appallingly discouraging statement repelled him away from this area, sending him in pursuit of fruitful research works elsewhere. Actually, the statement is an example of a common misinterpretation of the term ‘enhancement’ as applied to speech. Putting it more scientifically, what the researcher meant was, “Noise reduction and speech distortion are inversely proportional” [1]. Evidently, the researcher was referring to noise reduction (signal to noise ratio improvement), while mentioning speech enhancement. But because of this inverse relation, the more the noise is reduced the worse it sounds, increasing distortion and musical noise, regardless of all the sincere efforts.

Thus, in this research work, an attempt was undertaken to find a way to deal with both of these aspects of speech enhancement. It is well known that, despite their effectiveness in noise reduction, the commonly used spectral subtraction and Wiener filter based algorithms introduce distortion and annoying musical noise artifact in the enhanced speech. In this work our focus is to attain an improved noise reduction performance accompanied by a reduction of the musical in the enhanced speech using a two different stages. In the first stage, the focus is on noise reduction adapting a new MMSE estimator in DCT domain and in

²This is known as listener fatigue.

the second stage the focus is on reduction of the musical noise in the enhanced speech using a new non-stationary signal analysis tool known as the empirical mode decomposition [2].

1.2 Objective of this research

The purpose of this work is to develop a speech enhancement scheme with an improved noise reduction performance accompanied by a reduction of the musical noise in the enhanced speech. Since, we propose to attain these goals in two stages, the primary objective of this research is, therefore twofold.

1. Derivation of an optimum estimator for speech enhancement considering the constructive and destructive effects of noise.
2. Develop an effective residual noise removal method to be applied in the second stage.

The first objective is inspired by the fact that most traditional speech enhancement methods consider only the additive effect of noise, either inherently or effectively, giving an attenuating gain. However, in reality the noise coefficient may be both additive or subtractive with the clean signal. In this work, we consider both the constructive and destructive cases specifically and derive the optimum estimators in these given events. Thus, the new estimator will provide attenuation when the noise is constructive, but provide amplification when the noise is destructive. It will also handle the special case of polarity reversal which occurs when the noise coefficient is stronger than the signal coefficient. These situations were first taken into account in [3], assuming a Gaussian speech and noise statistical model. However, in their work, a linear MMSE estimator is assumed in the DCT domain resulting in a set of Wiener gains in the two cases. This inherently assumes that the joint distributions of noisy speech and clean speech in the two cases are jointly Gaussian. In this work, we assume the more accurate non-Gaussian joint probability distribution for noisy speech and clean speech in the two events and obtain a new non-linear MMSE estimator termed as the dual MMSE estimator (DMMSE). Our derivation results in a set of parametric gain curves that not only depend on the *a priori* information, but also utilize the instantaneous observation of the noisy coefficients.

Towards accomplishing the second objective of the work, we utilize the newly developed empirical mode decomposition (EMD). Very few applications of EMD applied to speech processing is found in the literature, which makes this an exciting field to work on. Here we propose a new post filtering technique for suppression of residual noise that remains in the enhanced speech utilizing the great versatility of EMD. We observed that EMD can effectively separate the residual noise of musical nature from the enhanced speech. This observation led us to utilize this decomposition for suppression of residual noise.

1.3 Organization of this thesis

This thesis consists of six chapters. Chapter 1 discusses the basics of speech enhancement, its importance in different applications and the main objective of this work. In Chapter 2, a brief review of the existing stochastic model based speech enhancement techniques is provided.

In Chapter 3, the dual MMSE estimator is derived. For the Gaussian statistical model, it is shown that the joint probability distribution in the conditional events, when the noise and speech coefficients are in constructive or destructive interference, are not jointly Gaussian. Using the accurate non-Gaussian joint probability density function, a new MMSE estimator is formed. Its properties and comparative performance is also discussed in this chapter.

In Chapter 4, the EMD based post processing technique is presented. The basics of the EMD algorithm and the properties of the IMFs are also discussed. Assuming a Chi-square probability density function for the short time IMF energy an optimum gain function is derived for residual noise suppression. The performance of the proposed post filtering method is tested by its application over traditional methods.

In Chapter 5, the performance of the hybrid speech enhancement method is compared to that of several speech enhancement techniques. Finally, some conclusions and suggestions for future works are provided in chapter 6.

Chapter 2

Stochastic model based speech enhancement: A review

2.1 Introduction

In high levels of ambient noise, the recorded speech picked up by any speech communication device becomes significantly impaired, reducing the quality and intelligibility of the transmitted speech signal. The degradations are usually very annoying, especially in mobile communications where hands free devices are often used in noisy environments. Also, this degradation causes a significant reduction of performance in speech recognition based systems that may be incorporated in the speech communications devices. Thus, speech enhancement is a crucial operation that must be performed as a pre-processing for these applications. However, there cannot be simply one “optimal” speech enhancement algorithm, mainly because of the large diversity of acoustic environments and noise reduction applications and their occasional demand of conflicting performance requirements. As a consequence, a variety of algorithms have been developed till today that have proved to be useful in either a certain noise environment or in a certain application.

Single channel speech enhancement techniques can be broadly classified into two categories: the ones based on stochastic models of speech and noise, and the others incorporating the perceptual aspects of speech signal and the auditory system. Each has its own limitations and advantages. The first category of algorithms are based on the variations of optimum filters and comprises such methods as spectral subtraction [4, 5], Wiener filtering [4, 6, 7], and various minimum mean square error (MMSE) spectral amplitude estimation methods

[8, 9, 10]. These algorithms are a common and effective way for enhancing speech degraded by acoustic additive noise given that only the noisy speech is available. The general requirements in this class of methods include: 1) a well defined suppression rule based on an optimality criteria [9, 6], which is usually a function of speech and noise statistics, 2) a method of estimating the speech and noise power spectral densities, 3) incorporation of the probability of speech presence to further attenuate non-speech bands [11], 4) a method for reducing residual noise by appropriately smoothing the estimated quantities [9, 12]. While, the first class of enhancement schemes perform optimization based on purely mathematical criteria, the second class of methods consider the auditory and perceptual criteria for performing enhancement. This class of enhancement system includes the perceptually-motivated processing such as critical-band filtering, lateral neural inhibition, and/or temporal/frequency masking [13, 14, 15, 16].

Since in this work, we are dealing with a new MMSE estimator in the DCT domain assuming a statistical model, only the relevant techniques will be reviewed in this chapter, with an emphasis on the attenuating and amplifying behavior of the resulting suppression rules. An elaborated review of statistical model based speech enhancement can be found in [17], while a comprehensive overview of speech enhancement techniques is available in [18, Chapter 8].

2.2 Short-time spectral subtraction

The spectral subtraction method, first proposed by Boll *et. al.* [4], is suitable for enhancing speech signals degraded by uncorrelated additive noise. It is an approach for estimating the power spectral density of the clean signal by subtracting an estimate of the power spectral density of the noise process from an estimate of the power spectral density of the degraded signal. The estimation is performed on a frame - by- frame basis, where each frame consists of 20 – 40ms of speech samples.

Let $\mathbf{x}[n]$, $\mathbf{d}[n]$ and $\mathbf{y}[n]$ denote vectors containing the L most recent samples of the clean signal, noise and noisy signal, respectively, in the i th analysis frame of size L . If it is assumed that the noise is additive, then,

$$\mathbf{y}[n] = \mathbf{x}[n] + \mathbf{d}[n]. \quad (2.1)$$

In the spectral subtraction method, the short-time spectral magnitude of the clean speech is estimated from (2.1) as,

$$|\hat{X}(i, k)|^2 = |Y(i, k)|^2 - E\{|D(i, k)|^2\} \quad (2.2)$$

where $Y(i, k)$ and $D(i, k)$ represents the discrete Fourier transforms (DFT) of $\mathbf{y}[n]$ and $\mathbf{d}[n]$ and k and i indicates the frequency and frame index, respectively. It is understood that a half wave rectification operation has to be performed on the right hand side of (2.2) since power spectral density cannot be negative. Since $|D(i, k)|^2$ is not directly available, it is approximated as $E\{|D(i, k)|^2\}$, where $E\{\cdot\}$ denotes the expectation operation. $E\{|D(i, k)|^2\}$ is obtained either from the assumed known properties of $\mathbf{d}[n]$ or by actual measurement during an interval when speech is absent. The spectral subtraction approach can be generalized by,

$$|\hat{X}(i, k)| = ||Y(i, k)|^\alpha - \beta E\{|D(i, k)|^\alpha\}|^{\frac{1}{\alpha}} \quad (2.3)$$

where constants α and β represent extra degrees of freedom used to enhance the algorithm's performance. Typical values are $\alpha = 2$ and $\beta = 1$ [5]. The estimate of clean speech segment $\hat{\mathbf{x}}[n]$ is obtained by combining $|\hat{X}(i, k)|$ with the phase of degraded signal $\angle Y(i, k)$ and then performing the inverse Fourier transform¹. In other words,

$$\hat{\mathbf{x}}[n] = \mathcal{F}^{-1}[|\hat{X}(i, k)| \cdot e^{j\angle Y(i, k)}], \quad (2.4)$$

where \mathcal{F}^{-1} denote the inverse Fourier transformation. The concept of the spectral subtraction method is based on the general idea that the additive noise has increased the signal transform coefficient, thus a subtraction is needed to be performed. Thus the method is inherently assuming that the noise coefficient was in a constructive interference with the signal. Accordingly, the method is generally effective at reducing the apparent noise power followed by an improvement in SNR. However, this noise reduction is achieved at the price of reduced speech intelligibility. A moderate amount of noise reduction can be achieved without significant intelligibility loss; however, a large amount of noise reduction can seriously degrade the intelligibility of the speech. Another disadvantage of the spectral subtraction approach is that they produce very annoying musical tones

¹It should be noted that the spectral subtraction method can also be easily implemented in the discrete cosine transform (DCT) domain. Only the polarity of the noisy DCT is to be combined with the clean estimate rather than the complex phase in case of DFT coefficients.

in the enhanced speech [19]. It is known that the musical noise occurs due to the random appearance and disappearance of spurious harmonics in the enhanced spectrum, since the method relies on subtracting an overall average noise spectrum. In reality the noise spectral variance is not a constant over the different frames and its effect on the signal spectral component is not always additive.

2.3 Wiener filtering and its variants

The Wiener filter is an algorithm that minimizes the expected error between the estimated speech and the actual speech signal assuming a multiplicative gain in the frequency domain, or a convoluting filter in the sample domain. In the frequency domain, it can be viewed as a an MMSE estimator that assumes a linear relation between the noisy coefficient and the estimated coefficient. If X_k and Y_k denote the clean speech and noisy speech transform coefficient in the k th bin, the goal of the Wiener filter is to find an estimate of the clean speech coefficient \hat{X}_k , such that

1. \hat{X}_k and Y_k are related by a multiplicative constant.
2. $E\{(\hat{X}_k - X_k)^2\}$ is minimum.

Usually, a multiplicative gain is assumed for the estimation, i.e. a linear equation such as, $\hat{X}_k = WY_k$ is assumed. If signal and noise samples are assumed to be uncorrelated stationary random processes with power spectral densities $P_x(k)$ and $P_d(k)$, respectively, the Wiener estimator for $\hat{x}(n)$ is found to be,

$$\hat{X}_k = \frac{P_x(k)}{P_x(k) + P_d(k)} Y_k. \quad (2.5)$$

This is known as the non-causal Wiener filter. However, speech cannot be assumed to be stationary. Which implies that the noncausal Wiener filter cannot be applied directly on all the speech samples. An approximation to noncausal Wiener filtering can be found if the gain is applied frame by frame. It is given by [19],

$$\hat{X}(i, k) = \frac{\hat{P}_x(i, k)}{\hat{P}_x(i, k) + \hat{P}_d(i, k)} Y(i, k) \quad (2.6)$$

where $\hat{P}_x(i, k)$ and $\hat{P}_d(i, k)$ are estimates for the short-term power spectrum for speech and noise, respectively, in the i -th frame and frequency bin k . Estimates

for the short-term speech power spectrum $\hat{P}_x(i, k)$ are obtained recursively. Estimates for the short-term speech power spectrum $\hat{P}_x(i, k)$ can also be obtained assuming an all-pole model for speech and forming a maximum a posteriori (MAP) estimate of the all-pole model parameters [20]. Since the iterative Wiener filtering approach was found to produce unnatural sounding speech and processing artifacts, certain constraints can be applied across iterations and across temporal frames to ensure the enhanced speech power spectrum has speech-like characteristics and remains mathematically stable [21].

Unlike the spectral subtraction approach, the Wiener filter does not explicitly assume that the noise power is additive in each spectral component of the noisy speech frame. However, since power spectral densities are positive quantities, (2.6) produce a multiplicative gain that is always attenuating, inherently assuming that the noise was additive in the spectral domain. This is due to the fact that, for additive uncorrelated noise, the noisy signal power is always greater than the clean speech power, demanding an attenuation for noise reduction.

2.3.1 Wiener filter in DCT domain

The wiener filter can be derived directly in the frequency domain. We shall discuss the DCT domain which is more relevant to this thesis work. If X_k , D_k and Y_k denote the clean speech, noise and noisy speech DCT coefficient in the k th bin and i th frame, for additive noise we have,

$$Y_k = X_k + D_k. \quad (2.7)$$

The goal of the Wiener filter is to minimize the cost function

$$J_W = E\{(\hat{X}_k - X_k)^2\}, \quad (2.8)$$

in MMSE sense. However, before minimizing (2.8), the Wiener filter assumes a multiplicative gain² for the estimated DCT coefficient \hat{X}_k . If G_W is the filter gain, an estimate of the clean speech spectral component is obtained as

$$\hat{X}_k = G_W Y_k. \quad (2.9)$$

²This simplification, reduces mathematical complexity and the difficulties of calculating the joint probability densities that are involved in the expectation operation in (2.8). In the next section we shall discuss the MMSE estimation without this assumption.

Substituting Eq. (2.9) into Eq. (2.8)

$$J_W = E\{(G_W Y_k - X_k)^2\}. \quad (2.10)$$

Substituting Eq. (2.7) into Eq. (2.10)

$$\begin{aligned} J_W &= E\{[G_W(X_k + D_k) - X_k]^2\} \\ &= (G_W^2 - 2G_W + 1)E\{X_k^2\} + 2G_W(G_W - 1)E\{X_k D_k\} \\ &\quad + G_W^2 E\{D_k^2\} \end{aligned} \quad (2.11)$$

Using the fact that X_k and D_k are real, zero-mean and uncorrelated random variables (i.e., $E\{X_k D_k\} = 0$, $E\{X_k\} = 0$ and $E\{D_k\} = 0$), the above cost function takes the form

$$J_W = (G_W^2 - 2G_W + 1)E\{X_k^2\} + G_W^2 E\{D_k^2\} \quad (2.12)$$

Differentiating J_W with respect to G_W gives

$$\begin{aligned} \frac{\partial J_W}{\partial G_W} &= (2G_W - 2)E\{X_k^2\} + 2G_W E\{D_k^2\} \\ &= 2(G_W - 1)E\{X_k^2\} + 2G_W E\{D_k^2\} \end{aligned} \quad (2.13)$$

Equating $\partial J_W / \partial G_W$ to zero yields

$$2(G_W - 1)E\{X_k^2\} + 2G_W E\{D_k^2\} = 0 \quad (2.14)$$

This leads to the optimum Wiener gain,

$$G_W = \frac{E\{X_k^2\}}{E\{X_k^2\} + E\{D_k^2\}}. \quad (2.15)$$

Note the similarity of the gain function with (2.6). The Wiener gain can be expressed in a much compact form as,

$$G_W = \frac{\xi_k}{1 + \xi_k}, \quad (2.16)$$

where

$$\xi_k = \frac{E\{X_k^2\}}{E\{D_k^2\}}. \quad (2.17)$$

is interpreted as the *a priori* SNR after McAullay and Malpass [7]. The value of ξ_k is calculated by using the decision directed method given in [9]. Equation (2.16) obviously gives a gain value that is always less than unity. Thus, as mentioned above, the conventional Wiener filter inherently assumes that the noise was in a constructive interference with the signal in the DCT domain.

2.3.2 The dual gain Wiener filter

The constructive and destructive interference of signal and noise DCT coefficients was first analyzed and incorporated in speech enhancement in [3]. The authors derived a set of multiplicative in the DCT domain assuming a two state model for the constructive and destructive interference of noise with the clean signal. A sign estimation algorithm was proposed to identify the constructive and destructive interference and the appropriate gain function was chosen to deal with the two events. Since this approach deals with the two cases separately and provides two different gains, it is termed as a dual gain filter. It will be apparent shortly that the multiplicative gains proposed in [3] are only Wiener gains for the conditional events. Thus, we denote these gains as the dual gain Wiener filter (DGW) in this thesis.

The development of the gains is directly related to the conventional Wiener derivation in the previous section. As for the Wiener filter derivation, the Gaussian distribution is assumed for the speech and noise DCT coefficients. However, in this case the Wiener filters are derived in two conditional events, that are mutually exclusive. The events are defined as,

$$H_+: \quad \text{signal and noise are constructive: } X_k D_k \geq 0,$$

$$H_-: \quad \text{signal and noise are destructive: } X_k D_k < 0.$$

If it is known that given one of these events have occurred, a pair of conditional MSE are obtained leading to two different multiplicative gains [3] in the two events. Denoting them as the dual gain Wiener, we have³

$$G_{W+} = \frac{\xi_k + \frac{2}{\pi}\sqrt{\xi_k}}{\xi_k + 1 + \frac{4}{\pi}\sqrt{\xi_k}}, \quad (2.18)$$

$$G_{W-} = \frac{\xi_k - \frac{2}{\pi}\sqrt{\xi_k}}{\xi_k + 1 - \frac{4}{\pi}\sqrt{\xi_k}}. \quad (2.19)$$

The gains G_{W+} and G_{W-} are to be used for constructive and destructive interference, respectively. Thus, the dual gain Wiener (DGW) estimator is given by

$$\hat{X}_k^{DGW} = [p_k G_{W+} + (1 - p_k) G_{W-}] \times Y_k. \quad (2.20)$$

³The derivation of these gains are given in the Appendix on section A.1

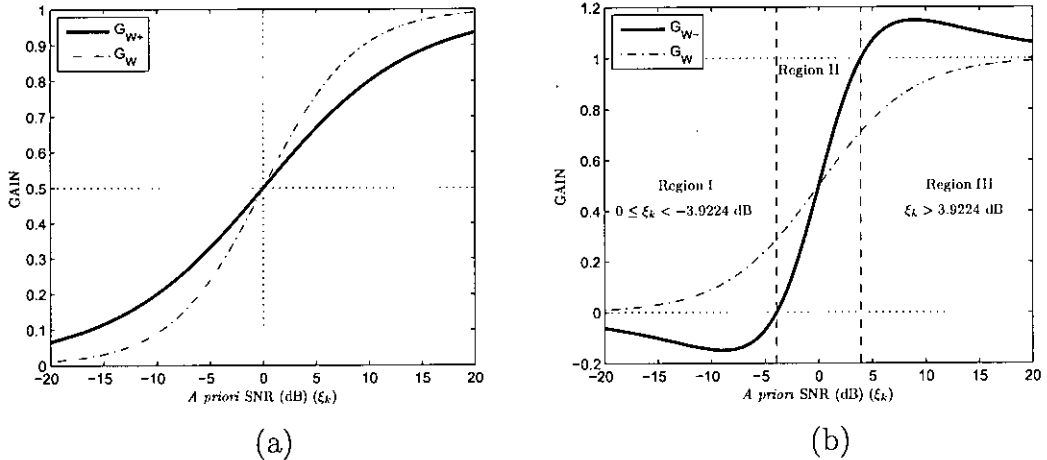


Fig. 2.1: Plots of the the dual gain Wiener (DGW) filters G_{W+} and G_{W-} proposed by [3] as given in (2.18) and (2.19). The gains are plotted against the variation of the *a priori* SNR values. The conventional Wiener gain G_W is also shown in the same plot.

where,

$$p_k = \begin{cases} 1 & \text{If event } H_+ \text{ is detected} \\ 0 & \text{If event } H_- \text{ is detected} \end{cases} \quad (2.21)$$

Both the gains G_{W+} and G_{W-} show interesting properties that are intuitively meaningful. Being the attenuating gain, G_{W+} is always less than unity, which is easily understood from (2.18). This is very desirable since in the event H_+ , D_k always serves to increase X_k . The gain G_{W+} , is more attenuating compared to the Wiener filter (G_W) when $\xi_k > 0$ dB and vice versa. At $\xi_k = 0$ dB both gains equal to $\frac{1}{2}$.

It is claimed in [3] that the gain G_{W-} is always greater than unity, serving to amplify the noisy coefficient Y_k , when noise is destructive. However, if we plot the gain curves with respect to the *a priori* SNR, we can clearly see that the gain G_{W-} is not always greater than unity as it should be. As a matter of fact, it even gives negative values at certain ranges. This befuddling and seemingly counter-intuitive phenomenon not discussed in [3], can only be explained if the polarity reversal case occurring in the destructive interference is taken into account.

The gain curve of G_{W-} can actually be divided into three separate regions as shown in Fig 2.1 (b). In region I, where the *a priori* SNR is very low, the gain is negative, that serves to encounter the polarity reversal caused by the noise when $|D_k| > |X_k|$. As ξ_k increases, the gain enters region II, giving an attenuation. This

is due to the uncertainty between a polarity reversal and magnitude reduction that might have occurred for the given Y_k . In this region, even if a polarity reversal occurs, the attenuation will reduce the error between the clean estimate at least to some extent. As ξ_k increases, the gain increases and finally gives amplification in region III, when the estimator assumes that $|X_k| > |D_k|$ resulting in a magnitude reduction in Y_k .

Thus, the authors of [3], only mentioned the region III of the gain curve which gives amplification, not mentioning the other regions. Moreover, the same gains were also applied in discrete Fourier transform (DFT) domain, which will provide meaningless results if the gain becomes negative. However, the main limitation of the dual gain Wiener filter is that it does not give the optimal solution in the assumed statistical model. It assumes a linear MMSE estimator in the DCT domain, which is the optimal MMSE only when the processes involved are jointly Gaussian. Chapter 3 of this thesis deals with the accurate modeling of these two events which are actually non-Gaussian.

2.4 Minimum mean square error (MMSE) estimators

The MMSE estimator is very similar to the Wiener filter, except that it does not assume a linear relation between the clean and noisy signal transform coefficients in the frequency domain. For this reason, the Wiener filter is known as a linear MMSE estimator. If X_k and Y_k denote the clean speech and noisy speech transform coefficient in the k th bin, the goal of the MMSE estimator is to find an estimate of the clean speech coefficient \hat{X}_k , such that the mean square error (MSE) given by,

$$\epsilon = E\{(\hat{X}_k - X_k)^2\} \quad (2.22)$$

is minimized. It is clear that, Y_k is the observed parameter and thus \hat{X}_k must be a function of Y_k . Now, unlike the Wiener estimator, no assumption will be made about the relation between \hat{X}_k and Y_k . Thus, (2.22) must be evaluated and minimized by solving the expected value. Since the expectation operator is on a function of X_k and Y_k , from (2.22),

$$\epsilon = \int \int (\hat{X}_k - X_k)^2 p(X_k, Y_k) dX_k dY_k \quad (2.23)$$

where, $p(X_k, Y_k)$ is the joint probability density function of X_k and Y_k . The limit of integration will depend on the type of transform coefficient used⁴. From (2.24) the distinction between the Wiener estimator and an MMSE estimator is very clear. Now, from (2.24) using the theory of conditional probability, we may write

$$\epsilon = \int p(Y_k) \int (\hat{X}_k - X_k)^2 p(X_k|Y_k) dX_k dY_k$$

Now, we have

$$E\{(\hat{X}_k - X_k)^2|Y_k\} = \int (\hat{X}_k - X_k)^2 p(X_k|Y_k) dX_k. \quad (2.24)$$

Thus, substituting (2.24) in (2.24), we have,

$$\epsilon = \int p(Y_k) E\{(\hat{X}_k - X_k)^2|Y_k\} dY_k \quad (2.25)$$

The density $p(Y_k)$ is nonnegative. Thus, to minimize the mean-squared error, it is sufficient to minimize the conditional expectation, $E\{(\hat{X}_k - X_k)^2|Y_k\}$ for each value of Y_k . It is well known from probability theory that, for a random variable x , $E\{(x - c)^2\}$ is minimum when c is the mean of the random variable x , i.e. $c = E\{x\}$. Thus, to minimize the conditional expectation $E\{(\hat{X}_k - X_k)^2|Y_k\}$, \hat{X}_k must be equal to the conditional mean. Which means, the MMSE estimator will be given by,

$$\hat{X}_k^{\text{MMSE}} = E\{X_k|Y_k\} \quad (2.26)$$

$$= \int \int X_k p(X_k|Y_k) dX_k \quad (2.27)$$

Thus, to find the MMSE estimator for the speech enhancement problem, this equation has to be solved. It is clear that the solution will involve the conditional probability density functions, and consequently the joint density functions of X_k and Y_k .

If the joint density function of X_k and Y_k is jointly Gaussian, that is, X_k and Y_k is assumed to be normally distributed, the MMSE solution and Wiener solution give the same results⁵[23]. Thus, depending on the assumed distribution of X_k and Y_k , the complexity of the MMSE estimation problem will vary.

⁴For absolute value of DFT coefficients, a limit of 0 to ∞ and for DCT coefficients 0 to ∞ would be used.

⁵This is shown in the MMSE estimator for DCT in the Appendix A.2.

2.4.1 MMSE estimator in the DFT domain

Assuming the Gaussian Model, Ephraim and Malah [9] derived a minimum mean-square error (MMSE) short-time spectral amplitude estimator under the assumption that the Fourier expansion coefficients of the original signal and the noise may be modeled as statistically independent, zero-mean, Gaussian random variables. Thus the observed spectral component in the bin k , $\mathbf{Y}_k \triangleq R_k \exp(j\vartheta_k)$, is equal to the sum of the spectral components of the signal, $\mathbf{X}_k \triangleq A_k \exp(j\alpha_k)$, and the noise, \mathbf{D}_k . That is,

$$\mathbf{Y}_k = \mathbf{X}_k + \mathbf{D}_k \quad (2.28)$$

Assuming this model, the MMSE estimator for the short time amplitude of the clean speech is shown to be [9]⁶,

$$\hat{A}_k = \Gamma(1.5) \exp\left(-\frac{\nu_k}{2}\right) \left[(1 + \nu_k) I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right) \right] R_k \quad (2.29)$$

where $\Gamma(\cdot)$ is the Gamma function, and $\Phi(a, b; z)$ is the confluent hypergeometric series defined in [22] and

$$\nu_k \triangleq \frac{\xi_k}{1 + \xi_k} \quad (2.30)$$

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \quad (2.31)$$

$$\gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)} \quad (2.32)$$

Thus the gain function is given by,

$$G_{\text{MMSE}}(\xi_k, \gamma_k) = \Gamma(1.5) \exp\left(-\frac{\nu_k}{2}\right) \left[(1 + \nu_k) I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right) \right] \quad (2.33)$$

This gain function will be termed as the Ephraim-Malah suppression rule (EMSR) in this thesis. It is a parametric gain function, since it depends not only on the *a priori* SNR, ξ_k , but also the *a posteriori* SNR, γ_k . A plot of this gain function is shown in Fig. 2.2.

2.4.2 MMSE estimator in the DCT domain

Assuming the DCT coefficients of the clean signal and noisy signal X_k and D_k , respectively, are Gaussian distributed random variables, the MMSE estimator \hat{X}_k

⁶The proof of this suppression rule is given in Appendix A.4.

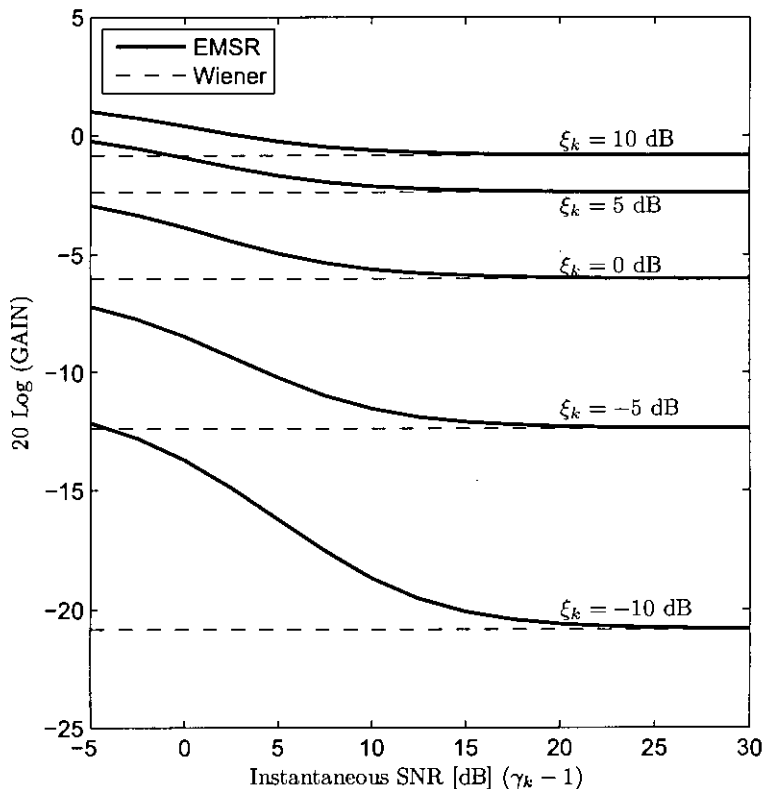


Fig. 2.2: The parametric gain curves of the Empraim-Malah suppression rule (EMSR).

simply gives the Wiener solution. That is,

$$\hat{X}_k^{\text{MMSE}} = \frac{\xi_k}{1 + \xi_k} Y_k \quad (2.34)$$

As stated before, this is expected since DCT coefficients are real and correspondingly, the noise and speech coefficients are jointly Gaussian [23]. This issue is also discussed in [24] and the proof is given in Appendix A.2.

2.4.3 Conclusion

A brief summary of statistical model based speech enhancement methods are discussed in this chapter. Emphasis was given on the dual gain Wiener filter, which is the only known method that considers the constructive and destructive effects of signal and noise transform coefficients separately. However, it also has some limitations. The classical MMSE estimators are also discussed. In the next chapter we shall proceed towards the proposed dual MMSE estimator.

Chapter 3

The proposed dual MMSE estimator

Nearly all orthodox speech enhancement algorithms use an attenuating filter in the transform domain for noise suppression, which inherently assumes that the noise transform coefficient is additive. But in reality, when both signal and noise exist in the same transform coefficient, the observed noisy coefficient magnitude may not always be greater than the clean signal even though the noise was originally additive.

In case of discrete Fourier transformation (DFT), the addition of complex signal and noise coefficients may or may not provide a resultant magnitude greater than the clean signal coefficient magnitude, depending on the relative phase angle between the signal and noise coefficients. Similarly, in the DCT domain, depending on the sign of the signal and noise coefficient, they can be either constructive or destructive. Since only the spectral amplitudes are estimated in conventional DFT based methods [9], [19], (keeping the phase angle of the noisy coefficient intact), the suppression rule should give attenuation and amplification when the noise is constructive and destructive, respectively. However, in the DCT domain, the clean signal coefficient may increase or decrease in magnitude and even can reverse in polarity by the noise. Since in this domain, the clean signal coefficients are directly estimated from the noisy observations [3], [24] (rather than only an amplitude estimation), it is clear that only an attenuation filter cannot be optimum for handling these cases. A reduction and increase in magnitude of the noisy coefficients should be followed by an amplifying and attenuating gain, respectively. When a noisy coefficient is reversed in polarity, the gain should be

negative to correct both its sign and magnitude. This obviously contradicts the traditional view of a non-negative suppression rule [10] that primarily deals with spectral amplitude estimation alone.

The constructive and destructive interference of signal and noise transform coefficients was first analyzed and incorporated in speech enhancement in [3]. The authors derived a set of multiplicative and subtractive filters in the DCT domain assuming a two state model for the constructive and destructive interference of noise with the clean signal. A sign estimation algorithm was proposed to identify the constructive and destructive interference and the appropriate gain function was chosen to deal with the two events. This method is discussed in detail in Section 2.3.2, which is termed as the dual gain Wiener filter.

In this work, a set of MMSE estimators for the DCT domain speech enhancement, in the conditional events of a constructive and destructive interference of noise is proposed. The major difference of the proposed technique with that of [3] is no linear MMSE estimator is assumed for the formulation.¹ We show that the joint density function of the clean signal and noisy signal DCT coefficient in the two-state model is non-Gaussian and the MMSE estimator results in a parametric gain function similar to the ones reported in [9]. Performance comparison of the dual gain Wiener filter and the proposed MMSE estimator in terms of SNR and MSE improvement is presented to demonstrate the superiority of our approach. The contribution of this work is therefore an accurate modeling of the constructive and destructive events in the speech enhancement problem assuming the appropriate non-Gaussian distributions.

3.1 The dual MMSE estimator

Assuming the Gaussian statistical model for the clean and noisy signal DCT coefficients, we denote x_k and y_k as instances of the random processes X_k and Y_k . Thus, using (A.32), their joint probability density function will be [26],

$$p_{XY}(x_k, y_k) = \frac{1}{2\pi\sigma_d\sigma_x} \exp \left[-\frac{x_k^2}{2\sigma_x^2} - \frac{(y_k - x_k)^2}{2\sigma_d^2} \right], \quad (3.1)$$

¹It is well known that a linear MMSE estimator gives the Wiener solution [25].

where σ_x^2 and σ_d^2 are the signal and noise variances in the k th DCT coefficient, respectively. Now, if \hat{X}_k is the MMSE estimator of X_k , we have

$$\hat{X}_k = E\{X_k|Y_k\}. \quad (3.2)$$

Using (3.1), \hat{X}_k reduces to the conventional Wiener estimator \hat{X}_k^W , given in (A.44) [24]. This result is expected since $p_{XY}(x_k, y_k)$ is jointly Gaussian and in this case the optimum MMSE estimator is linear [23]. However, if we assume that a polarity estimator is available to detect the events H_+ and H_- , we may derive two MMSE estimators $E\{X_k|Y_k, H_+\}$ and $E\{X_k|Y_k, H_-\}$ in the conditional events. Recall that, the events H_+ and H_- are defined as,

$$\begin{aligned} H_+ &: \quad \text{signal and noise are constructive: } X_k D_k \geq 0, \\ H_- &: \quad \text{signal and noise are destructive: } X_k D_k < 0. \end{aligned}$$

Thus, a generalized dual MMSE (DMMSE) estimator can be formulated as

$$\hat{X}_k^{\text{DMMSE}} = p_k E\{X_k|Y_k, H_+\} + (1 - p_k) E\{X_k|Y_k, H_-\}, \quad (3.3)$$

where p_k is defined in (2.21). The dual gain Wiener actually assumes that $E\{X_k|Y_k, H_+\}$ and $E\{X_k|Y_k, H_-\}$ can be found by multiplying the noisy coefficient by a gain function. This is equivalent to assuming a linear MMSE estimator. Thus, assuming $E\{X_k|Y_k, H_+\} = G_{W+} Y_k$ and $E\{X_k|Y_k, H_-\} = G_{W-} Y_k$ in (3.3) would lead to the DGW as given in (2.20). Obviously, this is suboptimal in the assumed model because the joint density of the processes involved are not Gaussian, as will be apparent shortly. Thus, we attempt to derive the expressions of $E\{X_k|Y_k, H_+\}$ and $E\{X_k|Y_k, H_-\}$, assuming the appropriate joint probability distribution.

First, we note that, Y_k is constructed from the two mutually exclusive events H_+ and H_- . Thus, we may write

$$\begin{aligned} E\{X_k|Y_k, H_+\} &= \int_{-\infty}^{\infty} x_k p(x_k|Y_k, H_+) dx_k \\ &= \int_{-\infty}^{\infty} x_k \frac{p(x_k, Y_k, H_+)}{p(Y_k, H_+)} dx_k \end{aligned} \quad (3.4)$$

$$\begin{aligned} E\{X_k|Y_k, H_-\} &= \int_{-\infty}^{\infty} x_k p(x_k|Y_k, H_-) dx_k \\ &= \int_{-\infty}^{\infty} x_k \frac{p(x_k, Y_k, H_-)}{p(Y_k, H_-)} dx_k \end{aligned} \quad (3.5)$$

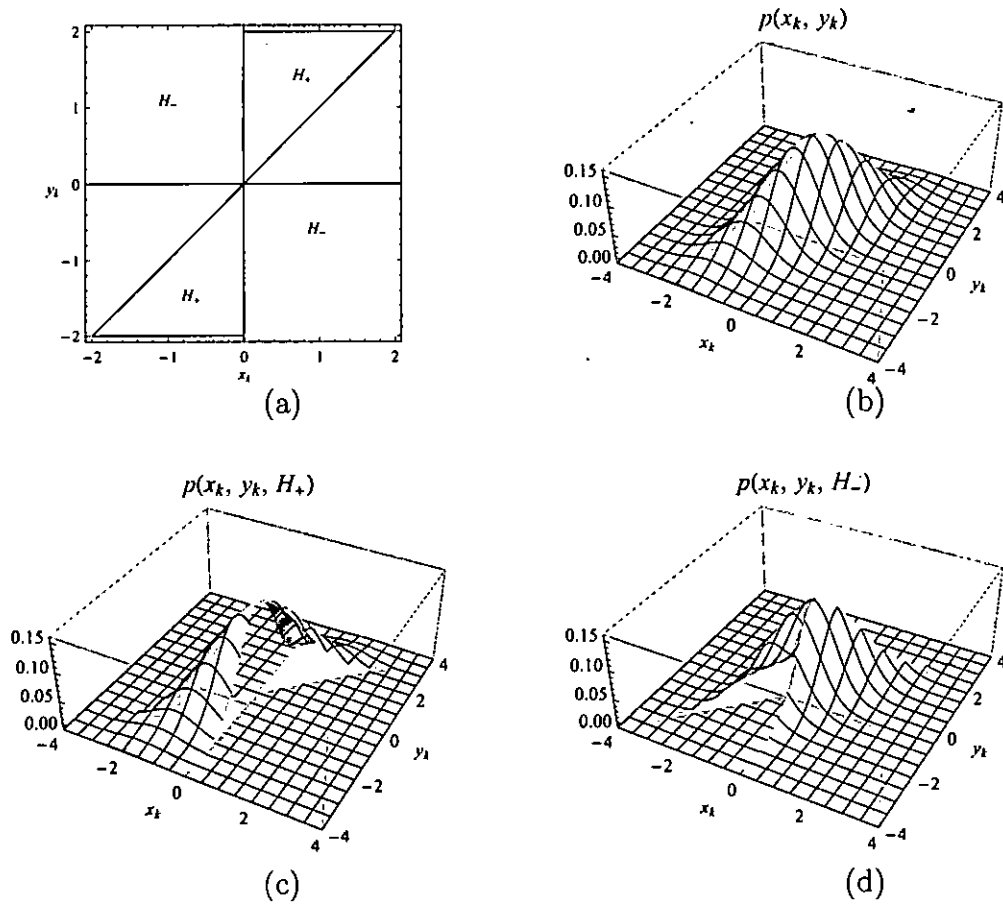


Fig. 3.1: (a) Valid regions of X_k and Y_k for the events H_+ and H_- , (b) The joint density function $p_{XY}(x_k, y_k)$, (c) The joint density function $p(x_k, y_k, H_+)$ and (d) The joint density function $p(x_k, y_k, H_-)$.

Thus the quantities in (A.46) and (A.47) can be determined if we know the joint density functions $p(x_k, y_k, H_+)$, $p(x_k, y_k, H_-)$, $p(Y_k, H_+)$ and $p(Y_k, H_-)$.

We note that, for the event H_+ , $X_k D_k > 0$, which results in $X_k Y_k > X_k^2$ using (A.32). This condition simplifies to $m_k Y_k > |X_k|$ where $m_k = \text{sgn}(X_k)$. Here $\text{sgn}(\cdot)$ is the signum function, given by,

$$\text{sgn}(\tau) = \begin{cases} +1 & \text{for } \tau > 0 \\ -1 & \text{for } \tau < 0 \\ 0 & \text{for } \tau = 0 \end{cases}$$

Similarly, the event H_- results in the condition $m_k Y_k < |X_k|$. These constraints of X_k and Y_k are shown graphically in Fig. 3.1 (a).

Thus the joint density functions $p(x_k, y_k, H_+)$ and $p(x_k, y_k, H_-)$ can be given

by,

$$p(x_k, y_k, H_+) = \begin{cases} p_{XY}(x_k, y_k) & m_k Y_k > |X_k| \\ 0 & \text{otherwise} \end{cases}$$

$$p(x_k, y_k, H_-) = \begin{cases} p_{XY}(x_k, y_k) & m_k Y_k < |X_k| \\ 0 & \text{otherwise} \end{cases}$$

These joint densities are plotted in Fig. 3.1(c) and Fig. 3.1(d) respectively, along with the joint density function $p_{XY}(x_k, y_k)$ in Fig. 3.1(b). As we can clearly see, these density functions in the conditional events H_+ and H_- are not jointly Gaussian due to their piecewise structure. Therefore, the optimum MMSE estimator in these conditional events will not be linear, as was assumed in [3].

To evaluate (A.46) and (A.47), we now need to find the probability densities $p(Y_k, H_+)$ and $p(Y_k, H_-)$. We have,

$$p(Y_k, H_+) = \int_{-\infty}^{\infty} p_{XY}(x_k, Y_k, H_+) dx_k.$$

The above integration is solved separately for positive and negative values of Y_k .

$$p(Y_k, H_+) = \begin{cases} \int_0^{Y_k} p_{XY}(x_k, Y_k) dx_k & \text{if } Y_k \geq 0 \\ \int_{Y_k}^0 p_{XY}(x_k, Y_k) dx_k & \text{if } Y_k < 0 \end{cases}$$

$$= \text{sgn}(Y_k) \int_0^{Y_k} p_{XY}(x_k, Y_k) dx_k. \quad (3.6)$$

Substituting the value of $p_{XY}(x_k, Y_k)$ from (3.1) into (A.52) and solving yields,

$$p(Y_k, H_+) = \text{sgn}(Y_k) \frac{e^{-\frac{Y_k^2}{2(\sigma_d^2 + \sigma_x^2)}} [\text{erf}(f_1) + \text{erf}(f_2)]}{2\sqrt{2\pi}\sqrt{\sigma_d^2 + \sigma_x^2}}, \quad (3.7)$$

where,

$$f_1 = \frac{Y_k \sigma_d}{\sqrt{2\sigma_x} \sqrt{\sigma_d^2 + \sigma_x^2}}$$

$$f_2 = \frac{Y_k \sigma_x}{\sqrt{2\sigma_d} \sqrt{\sigma_d^2 + \sigma_x^2}}$$

Similarly, $p(Y_k, H_-)$ is obtained as,

$$p(Y_k, H_-) = \text{sgn}(Y_k) \frac{e^{-\frac{Y_k^2}{2(\sigma_d^2 + \sigma_x^2)}} [\text{erfc}(f_1) + \text{erfc}(f_2)]}{2\sqrt{2\pi}\sqrt{\sigma_d^2 + \sigma_x^2}}. \quad (3.8)$$

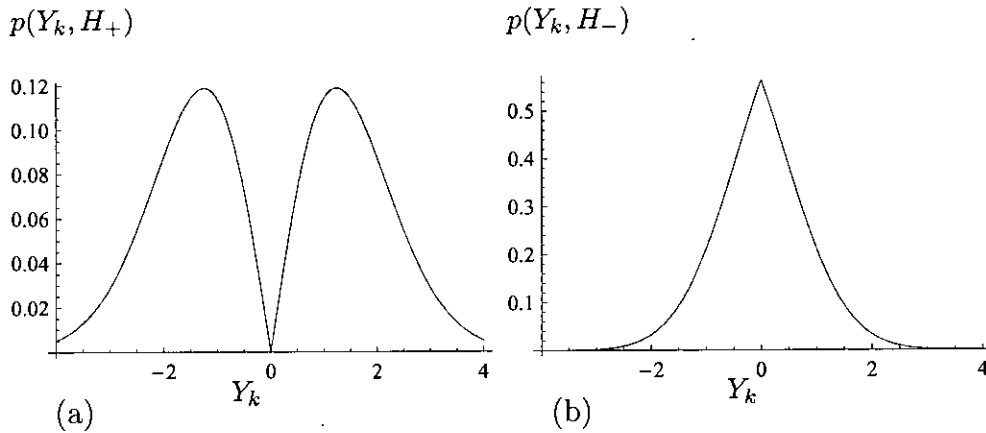


Fig. 3.2: (a) The probability density function $p(Y_k, H_+)$ and (b) the probability density function $p(Y_k, H_-)$ when $\sigma_d = \sigma_x = 1$.

The probability distributions $p(Y_k, H_+)$ and $p(Y_k, H_-)$ are shown in Fig. 3.2 for the case of $\sigma_x = \sigma_d = 1$. The shape of these density functions are quite interesting and intuitively meaningful as well. When the signal and noise are constructive, they are less likely to be near zero, which is the reason why the density in Fig. 3.2(a) has a notch shape at $Y_k = 0$. Again, when signal and noise are destructive, Y_k has a higher probability to be near zero as seen in Fig. 3.2(b). The sum of these two density functions obviously gives a Gaussian PDF.

Returning to (3.3), we now require the expressions $E\{X_k|Y_k, H_+\}$ and $E\{X_k|Y_k, H_-\}$ to determine the dual MMSE estimator. Approaching in a way similar to the previous derivation we obtain,

$$\begin{aligned}
 E\{X_k|Y_k, H_+\} &= \int_{-\infty}^{\infty} x_k p(x_k|Y_k, H_+) dx_k \\
 &= \frac{\int_{-\infty}^{\infty} x_k p(x_k, Y_k, H_+) dx_k}{p(Y_k, H_+)} \\
 &= \text{sgn}(Y_k) \frac{\int_0^{Y_k} x_k p_{XY}(x_k, Y_k) dx_k}{p(Y_k, H_+)}.
 \end{aligned}$$

Solving the integration yields,

$$E\{X_k|Y_k, H_+\} = G_W Y_k + \Phi(Y_k), \quad (3.9)$$

where,

$$G_W = \frac{\sigma_x^2}{\sigma_d^2 + \sigma_x^2},$$

$$\Phi(Y_k) = \sqrt{\frac{2}{\pi}} \left[\frac{\exp(-f_1^2) - \exp(-f_2^2)}{\operatorname{erf}(f_1) + \operatorname{erf}(f_2)} \right] \frac{\sigma_d \sigma_x}{\sqrt{\sigma_d^2 + \sigma_x^2}}. \quad (3.10)$$

Noting that $\Phi(Y_k)$, containing f_1 and f_2 , is an odd function, it can be expressed as

$$\Phi(Y_k) = \operatorname{sgn}(Y_k) \Phi(|Y_k|),$$

enabling us to express (A.59) as a gain expression multiplied by the noisy DCT component, Y_k , i.e.,

$$\begin{aligned} E\{X_k|Y_k, H_+\} &= \left(G_W + \frac{\Phi(Y_k)}{Y_k} \right) Y_k \\ &= \left(G_W + \frac{\Phi(|Y_k|)}{|Y_k|} \right) Y_k \\ &= G_{\text{MMSE}+} \times Y_k \end{aligned} \quad (3.11)$$

where $G_{\text{MMSE}+}$ denotes the MMSE gain function for the event H_+ . Expressing G_W and $\Phi(Y_k)$ using *a priori* and *a posteriori* SNR, the formulation of $G_{\text{MMSE}+}$ is given by,

$$G_{\text{MMSE}+} = \frac{\xi_k}{\xi_k + 1} + \sqrt{\frac{\xi_k}{1 + \xi_k}} \sqrt{\frac{2}{\pi \gamma_k}} \left[\frac{e^{-\frac{\gamma_k}{2\xi_k(1+\xi_k)}} - e^{-\frac{\gamma_k \xi_k}{2(1+\xi_k)}}}{\operatorname{erf}\left(\sqrt{\frac{\gamma_k}{2\xi_k(1+\xi_k)}}\right) + \operatorname{erf}\left(\sqrt{\frac{\gamma_k \xi_k}{2(1+\xi_k)}}\right)} \right], \quad (3.12)$$

where, $\gamma_k = \frac{|Y_k|^2}{E\{D_k^2\}}$, is termed as the *a posteriori* SNR after McAullay and Malpass [7]. Following a very similar method, the gain for the destructive case, $G_{\text{MMSE}-}$ can be expressed as,

$$G_{\text{MMSE}-} = \frac{\xi_k}{\xi_k + 1} - \sqrt{\frac{\xi_k}{1 + \xi_k}} \sqrt{\frac{2}{\pi \gamma_k}} \left[\frac{e^{-\frac{\gamma_k}{2\xi_k(1+\xi_k)}} - e^{-\frac{\gamma_k \xi_k}{2(1+\xi_k)}}}{\operatorname{erfc}\left(\sqrt{\frac{\gamma_k}{2\xi_k(1+\xi_k)}}\right) + \operatorname{erfc}\left(\sqrt{\frac{\gamma_k \xi_k}{2(1+\xi_k)}}\right)} \right] \quad (3.13)$$

Thus, the proposed dual MMSE estimator is given by

$$\hat{X}_k^{\text{DMMSE}} = [p_k G_{\text{MMSE}+} + (1 - p_k) G_{\text{MMSE}-}] \times Y_k. \quad (3.14)$$

In the following sections, the properties of the proposed MMSE gains $G_{\text{MMSE}+}$ and $G_{\text{MMSE}-}$ are discussed.

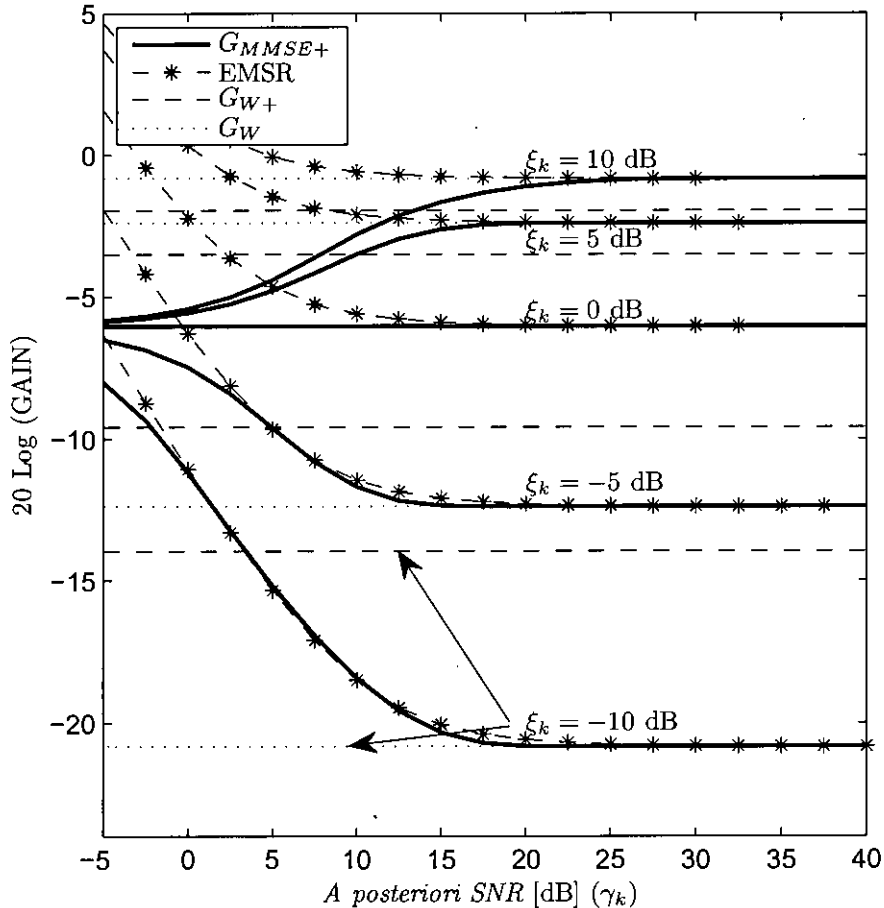


Fig. 3.3: Parametric gain curves describing (a) MMSE gain function in the constructive case, G_{MMSE+} (solid lines), (b) the gain function due to Ephraim-Malah (EMSR) (dash starred line), (c) the gain G_{W+} from the dual gain Wiener (dashed lines) as in (2.18) and (d) the Wiener gain function.

3.2 Properties of the gain G_{MMSE+}

The gain curves in Fig. 3.3 show the variation of gain G_{MMSE+} with the *a priori* SNR, ξ_k , and *a posteriori* SNR, γ_k . For comparison, we plot the well known gain function due to Ephraim and Malah [9] termed as EMSR (Ephraim-Malah Suppression Rule), the gain G_{W+} in (2.18) proposed in [3], and the conventional Wiener gain function G_W given in (A.45). The latter two gain functions only depend on ξ_k , whereas the former ones on both ξ_k and γ_k . Both these gains converge to the Wiener gain as the *a posteriori* SNR is increased provided that the *a priori* SNR is at a constant value. This is apparent from (3.12), since if

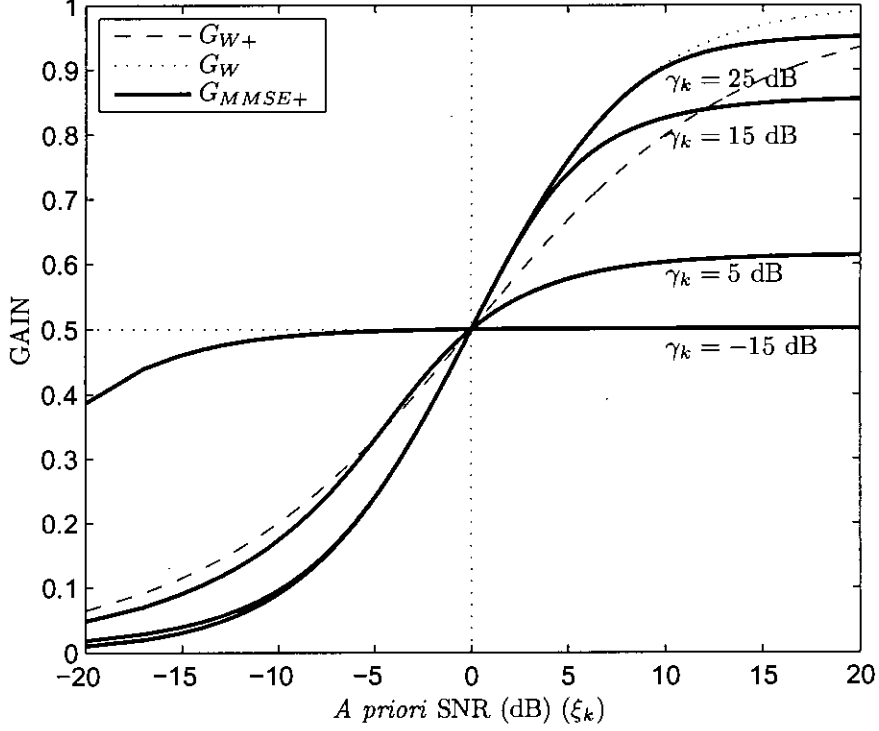


Fig. 3.4: Parametric gain curves plotted against ξ_k describing (a) MMSE gain function in the constructive case, $G_{\text{MMSE}+}$ (solid lines), (b) the gain G_{W+} from the dual gain Wiener (dashed lines) as in (2.18), and (c) the Wiener gain function (dotted lines).

one applies the limit $\gamma_k \rightarrow \infty$, the second term vanishes as

$$\lim_{\gamma_k \rightarrow \infty} G_{\text{MMSE}+} = \frac{\xi_k}{\xi_k + 1}$$

However, for lower values of γ_k , the two gains show spectacular divergence. When the *a posteriori* SNR tends to $-\infty$ dB, EMSR gain tends to ∞ dB, whereas $G_{\text{MMSE}+}$ tends to -6.02 dB, that is $\frac{1}{2}$. This can also be seen by taking the limits in (3.12).

$$\lim_{\gamma_k \rightarrow 0} G_{\text{MMSE}+} = \frac{1}{2}$$

Since the gain $G_{\text{MMSE}+}$ only deals with the constructive interference, it is always attenuating, i.e. below 0 dB as expected. However, the EMSR eventually goes beyond the 0 dB line for all the gain curves, when γ_k is very low, which means that EMSR inherently assumes the destructive interference and gives an amplifying gain for the weak noisy spectral components². Since EMSR is a spectral amplitude estimator, it can never give negative gain values, i.e.,

²It may be noted that EMSR is a spectral amplitude estimator for a complex Fourier ex-

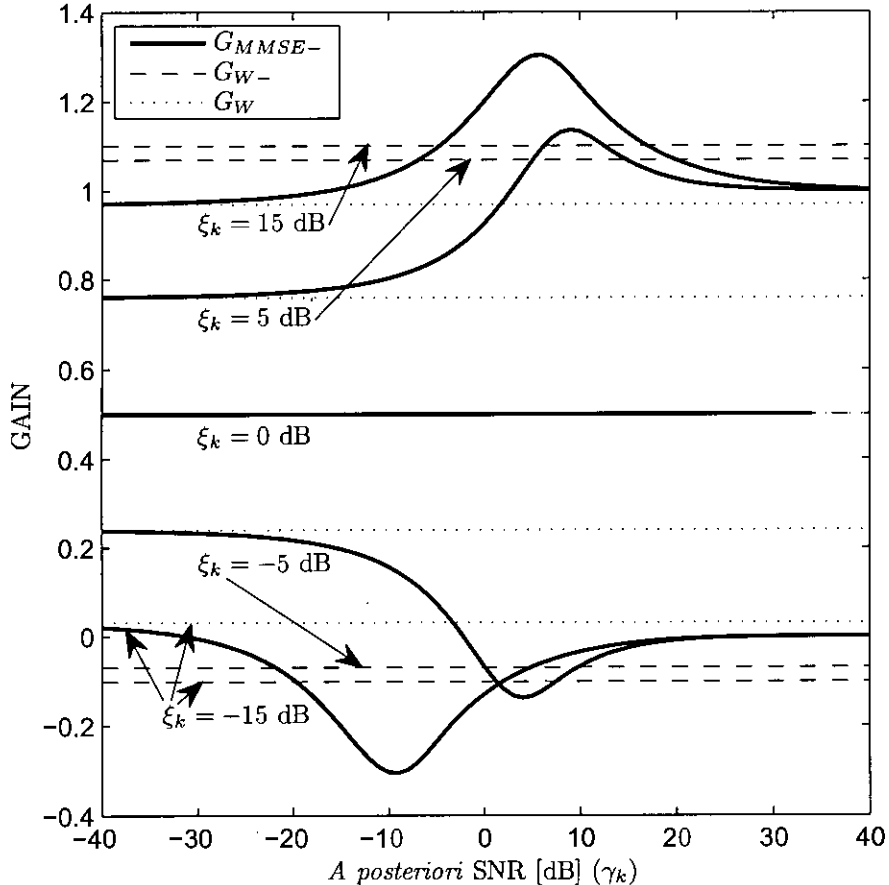


Fig. 3.5: Parametric gain curves describing (a) MMSE gain function in the destructive case, G_{MMSE-} (solid lines), (b) the gain G_{W-} from the dual gain Wiener as in (2.19); and (c) the Wiener gain function (dotted lines).

As the Wiener gain and the dual gain Wiener filters only depend on the *a priori* SNR, we plot the gains G_{MMSE+} , G_{W+} and G_W with the variation of ξ_k in Fig. 3.4 that exhibits the respective influence of γ_k and ξ_k [12]. These curves also demonstrate the convergence of the gain G_{MMSE+} to the Wiener gain and a constant $\frac{1}{2}$ when $\gamma_k \rightarrow \infty$ and $\gamma_k \rightarrow 0$, respectively.

3.3 Properties of the gain G_{MMSE-}

The gain curves in Fig. 3.5 shows the variation of the gain G_{MMSE-} with *a priori* SNR ξ_k , and *a posteriori* SNR, γ_k . Again, we compare the parametric gain curves with the Wiener (A.45) and the gain G_{W-} (2.19) proposed by [3]. We leave out

pansion coefficient. Thus the definition of the destructive interference is not the same as in case of DCT. In the context of complex Fourier coefficients, the destructive interference merely indicates that the absolute value of the noisy signal coefficient has been decreased by the noise.

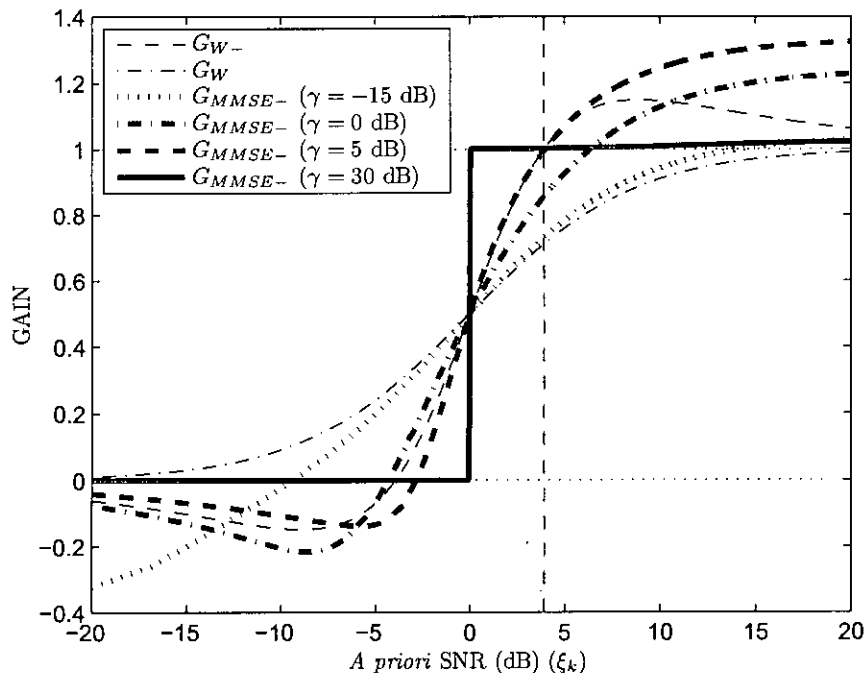


Fig. 3.6: Parametric gain curves plotted against ξ_k describing (a) MMSE gain function in the destructive case, $G_{\text{MMSE}-}$ (solid lines), (b) the gain G_{W-} from the dual gain Wiener as in (2.18), and (c) the Wiener gain function (dotted lines).

the EMSR in this case because our gains take on negative values which cannot be shown in log scale with the EMSR.

The gain $G_{\text{MMSE}-}$, in contrast to the EMSR and $G_{\text{MMSE}+}$, converges to the Wiener as the *a posteriori* SNR tends to $-\infty$ dB, i.e., 0. This can be shown mathematically from (3.13) as

$$\lim_{\gamma_k \rightarrow 0} G_{\text{MMSE}-} = \frac{\xi_k}{\xi_k + 1}$$

As γ_k increases, the convergence of the $G_{\text{MMSE}-}$ depends on ξ_k . If $\xi_k > 0$ dB, $G_{\text{MMSE}-}$ converges to 0 dB or unity gain. But if $\xi_k < 0$ dB, the gain converges to zero. For the case $\xi_k = 0$ dB, the gain is a constant at $\frac{1}{2}$ or -6.02 dB.

These situations are better observed when the gains are plotted against the variation of ξ_k as shown in Fig. 3.6. For the different values of γ_k the gain $G_{\text{MMSE}-}$ provides a gain curve similar to the DGW gain G_{W+} as shown in Fig. 2.1(b). All of these gain curves show three regions similar to G_{W-} . Interestingly, when γ_k tends to ∞ , the gain actually approaches a unit step function with unity lag, i.e.,

$$\lim_{\gamma_k \rightarrow \infty} G_{\text{MMSE}+} = U(\xi_k - 1),$$

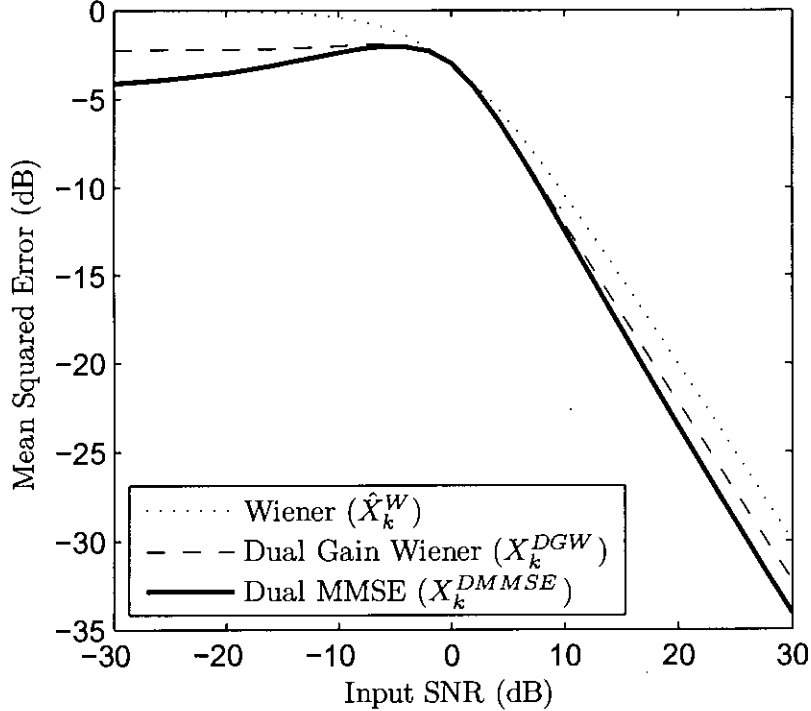


Fig. 3.7: Theoretical performance comparison of the conventional Wiener, dual gain Wiener and the proposed dual gains in the known polarity case.

where $U(\cdot)$ denotes the unit step function. This is evident from the $\gamma = 30$ dB gain curve shown in Fig. 3.6. This means, for very high spectral components, this MMSE estimator simply needs to decide if this high value was due to the signal or noise only, since it is given that the destructive interference has occurred. Thus, if ξ_k is greater than unity, the estimator decides that X_k was stronger and gives a unity gain and conversely if ξ_k is less than unity, it gives a zero gain assuming that D_k was stronger.

3.4 Performance comparison and discussion

3.4.1 Experiment using generated Gaussian sequences

In this section, we compare the performance of the conventional Wiener estimator (A.44), the dual gain Wiener estimator [3] as given in (2.20) and the proposed dual MMSE estimator (3.14) with respect to MSE value and SNR improvement. For this experiment, an ideal case is considered. Gaussian sequences of length $L = 40000$ are generated for signal $\mathbf{x}[n]$ and noise $\mathbf{d}[n]$ and are mixed in different

SNRs to construct the noisy signal $\mathbf{y}[n]$. The true polarities of the corresponding DCT coefficients X_k and D_k are assumed to be known, and the values of p_k set accordingly. The values of ξ_k are calculated from the known input SNR level. Using (A.44), (2.20) and (3.14) the clean estimates \hat{X}_k^W , \hat{X}_k^{DGW} and \hat{X}_k^{DMMSE} are calculated.

The improvement in MSE of the various estimators for different input SNR values is shown in Fig. 3.7. It is clear from the figure that the proposed dual MMSE estimator gives a better performance than those of the conventional Wiener and the dual gain Wiener estimators for all input SNR values. With respect to the DGW, the improvement is prominent for high and low values of the input SNRs. Note that the MSE values are equal for all the estimators when the input SNR is 0dB. This is expected since at 0dB SNR, all of the gains are equal to 0.5. This is seen in Figs. 3.4, 3.3, 3.6, 3.5. This theoretical performance comparison proves the effectiveness of the derived MMSE estimators for the assumed statistical model.

3.4.2 Experiments on speech files

The proposed dual MMSE estimator presented in this section, is tested using 5 male and 5 female utterances taken from the TIMIT database. The utterances are corrupted with white noise, taken from the 'NOISEX' database. The noise level is adjusted so as to give SNR from -10dB to 35dB. The sampling frequency is 8 KHz. A frame size of 32 ms (512 samples) is used for framing and the overlap-add method with 75% overlap is used for signal decomposition. 1024 point DCT is performed in each frame. The *a priori* SNR is calculated using the decision directed approach [9] utilizing different averaging parameter values.

As discussed in [3], a polarity estimation method is required to apply the dual gains in practical cases. However, in this work, the focus is on the optimal estimation of clean speech assuming an accurate modeling of the constructive and destructive interference. Thus, instead of actually using a separate polarity estimation algorithm, its operation is simulated assuming that the actual polarities of the noise DCT coefficients are known. The value of p_k are therefore computed as

$$p_k = U(X_k D_k) \quad \text{Where, } k = 1, 2, \dots, 512, \quad (3.15)$$

Results for $A_p = 100\%$ and $\alpha = 0.8$

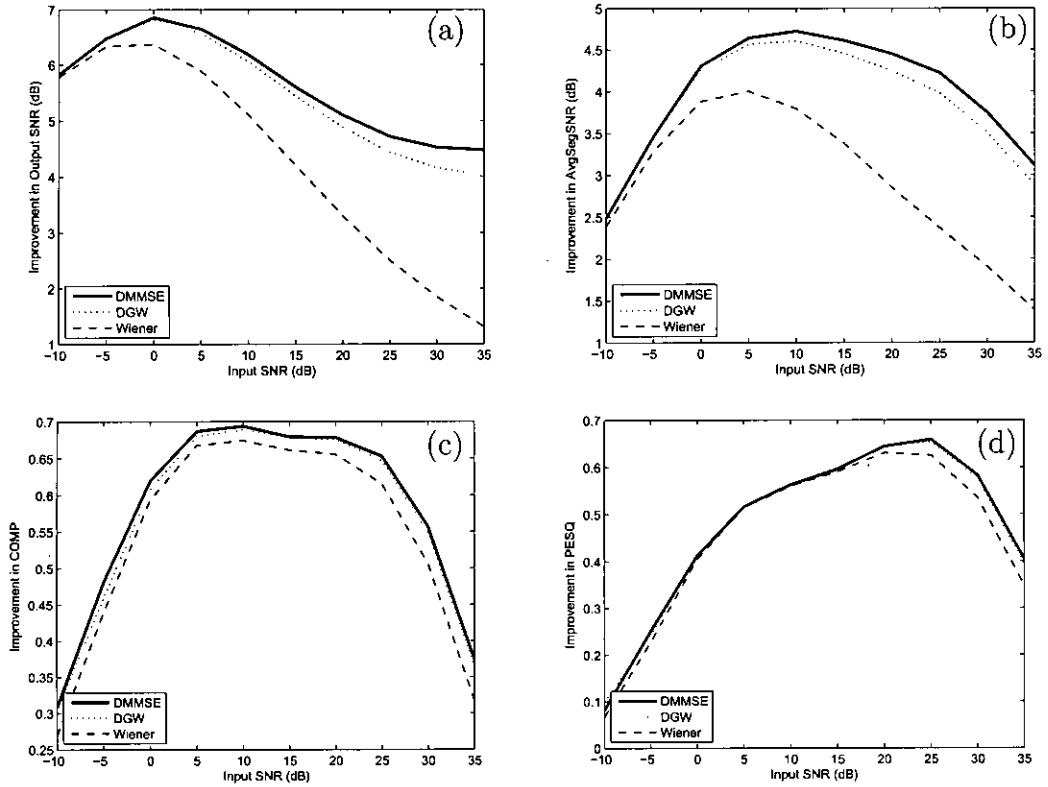


Fig. 3.8: Performance comparison of conventional Wiener (dashed lines), dual gain Wiener (DGW) (dotted lines) and the Proposed dual MMSE (DMMSE) estimator (solid lines) with respect to improvement in overall SNR, average segmental SNR, composite speech quality measure (COMP) and PESQ scores. Polarity estimator accuracy was assumed to be 100% and the averaging parameter α was set to 0.8.

This equation actually leads to (2.21). If an estimator could give the exact values of p_k , it would have had an accuracy of 100%. For simulating a polarity estimator having an accuracy of 80% (79.98% to be exact), 205 values of p_k are toggled (performing the NOT operation) randomly to find \hat{p}_k . This generated value, \hat{p}_k instead of p_k , is used in (2.20) and (3.14) the DGW and DMMSE estimators. The accuracy of the polarity estimator is denoted in the figures as A_p .

The results obtained from the conventional Wiener filter, the DGW and DMMSE estimator are presented for comparison. The averaged results of the 10 utterances are plotted in Figs. 3.8, 3.9 and 3.10. From these figures, it is clear that the DMMSE method shows uniform improvement with respect to output SNR, Average Segmental SNR, Composite speech quality measure (COMP) and

perceptual evaluation of speech quality (PESQ) scores for the complete range of input SNRs as compared to the Wiener and DGW estimators, when the averaging parameter $\alpha = 0.90$. For other values of α , an improvement in all ranges of input SNR is not always achieved. This indicates that even though the DMMSE method is consistently superior in the ideal case where real Gaussian sequences are used, in real case such as speech, the performance depends on the *a priori* SNR estimation.

It is to be noted that the gains are applied over a specific frequency bin across each time frames. Thus, the correlation between the successive frame DCT coefficients come into play. However, ideally this sample by sample correlation should not exist. This is why in the ideal case the DMMSE is consistently better than DGW. Since in reality, the successive frame DCT coefficients of the noisy speech is not Gaussian, the errors in estimation can be expected. It may be argued that why DGW would perform better than DMMSE in some cases given that DMMSE assumes the more accurate model. In this regard, our opinion is that, since DMMSE more rigorously considers the Gaussian model, it is more likely to give errors due to non-Gaussian speech DCT samples than the DGW estimator. DMMSE is more dependent on the Gaussian statistical model than the DGW. However, the ideal performance curve clearly demonstrates that DMMSE is superior to DGW if the data sequence really follows a Gaussian distribution.

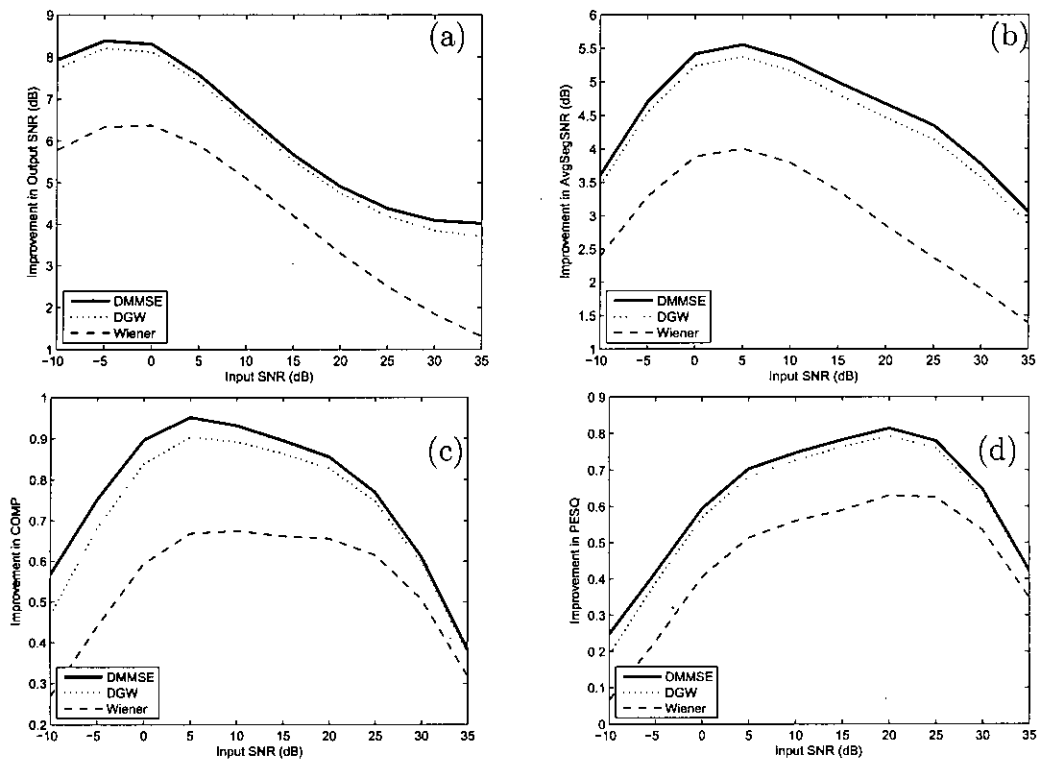
Both the DMMSE and DGW estimators show remarkable improvement over the Wiener estimator, which is actually demonstrates the effectiveness of the dual gain approach. With the polarity estimator accuracy of 100%, we have the highest improvement achievable theoretically. However, even for an accuracy of 80% the estimator performance is consistently superior to that of the DGW estimator as can be seen from Fig. 3.9.

3.5 Conclusion

A new MMSE estimator for DCT domain speech enhancement has been presented in this chapter. Unlike the traditional MMSE estimator approaches, we have deliberately considered the constructive and destructive interference between the signal and noise DCT coefficients. Similar to the previously proposed dual gain Wiener estimator, a dual MMSE estimator is formulated with the major

difference in the non-linear MMSE assumption. The proposed parametric gain characteristics are analyzed in detail and compared with the well known traditional gain functions. The theoretical performance of the proposed method has been evaluated using computer generated Gaussian sequences, showing a notable MSE reductions. The effectiveness of the proposed estimator has also been tested using speech files taken from the TIMIT database and significant performance improvement was achieved compared to the traditional estimators. In the following chapter, the proposed post-filtering method is described.

Results for $A_p = 100\%$ and $\alpha = 0.9$



Results for $A_p = 80\%$ and $\alpha = 0.9$

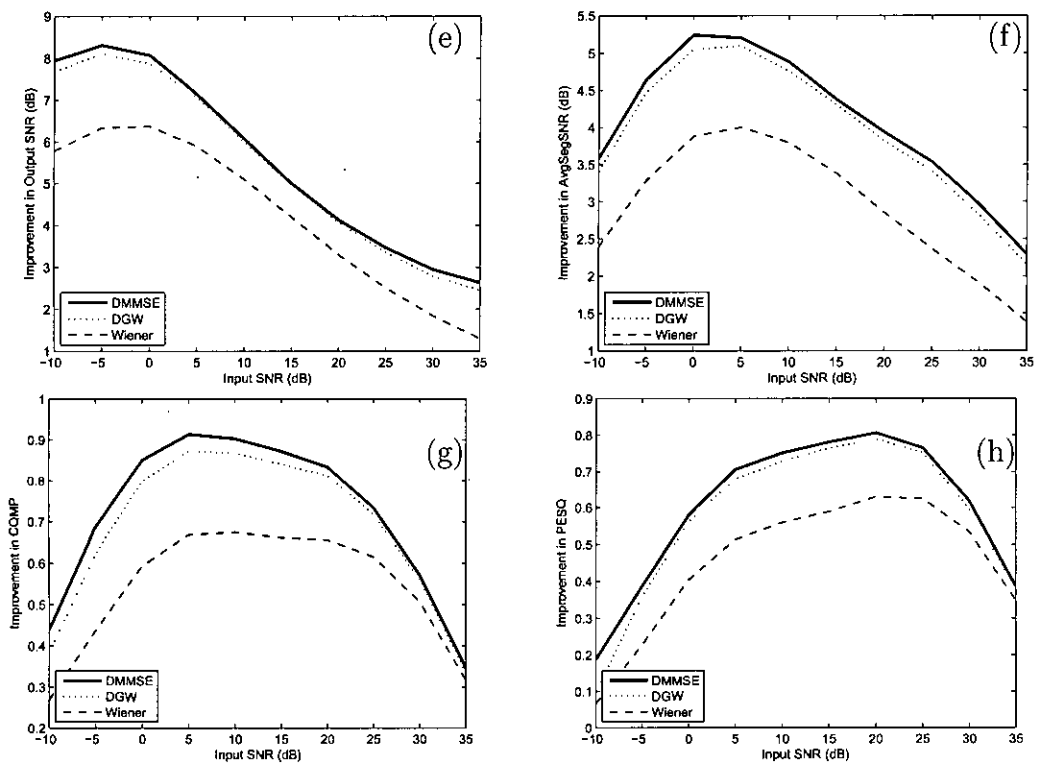
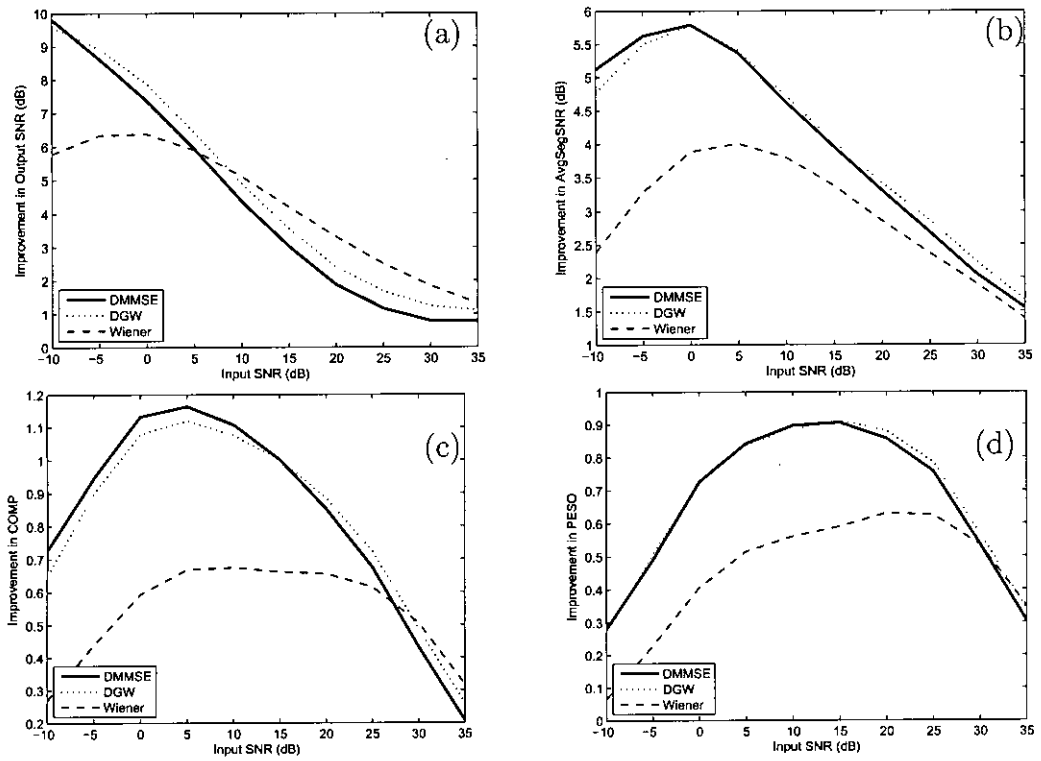


Fig. 3.9: Performance comparison of Wiener (---), DGW (····) and DMMSE estimator (—) for $\alpha = 0.9$. $A_p = 100\%$ for Figs. (a)-(d) and $A_p = 80\%$ for Figs. (e)-(h)

Results for $A_p = 100\%$ and $\alpha = 0.98$



Results for $A_p = 100\%$ and variable α [27]

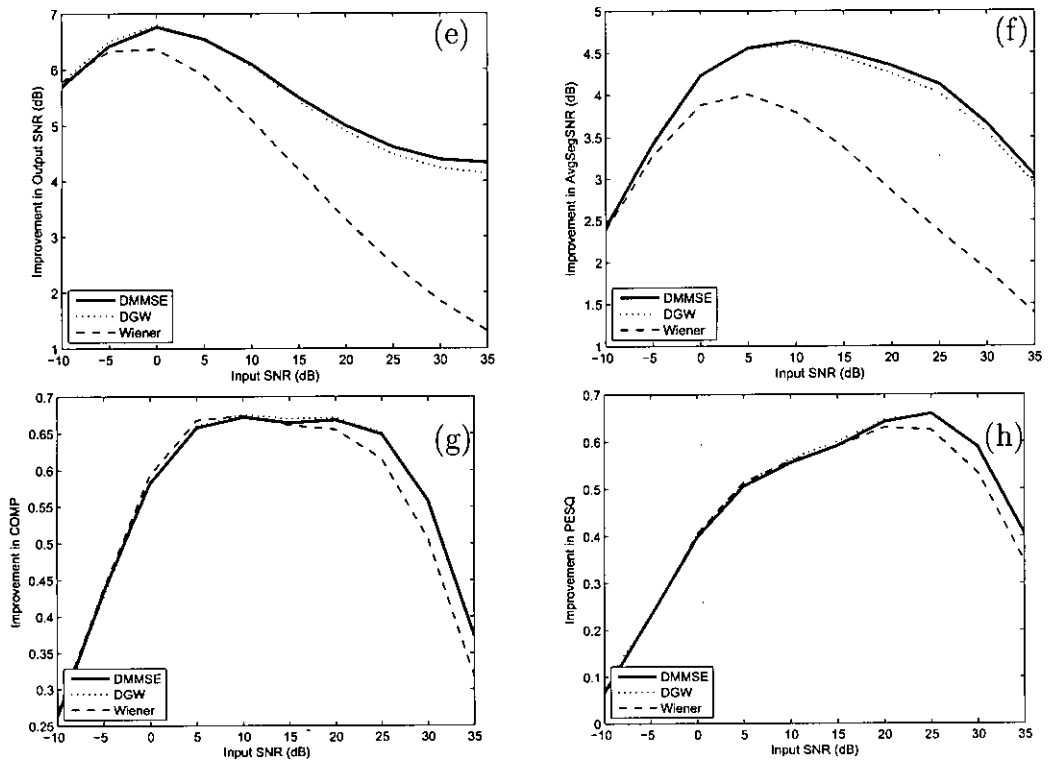


Fig. 3.10: Performance comparison of Wiener (---), DGW (····) and DMMSE estimator (—) for $A_p = 100\%$. $\alpha = 0.98$ for Figs. (a)-(d) and α is variable for Figs. (e)-(h)

Chapter 4

Post processing using the empirical mode decomposition

Most of the conventional speech enhancement methods can improve the signal to noise ratio (SNR) levels at the expense of introducing musical noise in the enhanced speech. A key element in this trade-off is the smoothing parameter α , used in the decision directed approach [9] for estimating the *a priori* SNR. As an example, the approximate MAP estimator [10] is well known to have very low residual noise, if $\alpha \sim 0.98$. But the SNR improvement of the method is not optimal for such a high value of α . A relatively lower value of it, or an adaptive α [27], gives much better SNR improvement, but also generates very annoying musical tones.

In this chapter, we show that the musical noise in the speech enhanced by popular methods, can be significantly suppressed using the newly developed empirical mode decomposition (EMD) [2], pioneered by Huang *et. al.* This noise suppression is accompanied by significant improvement of both subjective and objective quality measures.

While Flandrin *et. al.* gave a general discussion on de-noising using EMD [28], its application on speech enhancement is rather new. Among the few works, in [29] the authors have noted that the EMD method can be used to separate sudden impulsive unwanted sounds from speech. However, no formulation of noise power estimation, rigorous experiments and relevant quality index results were presented. In [30], a more analytical approach has been considered. The authors discussed the noise power densities in each intrinsic mode functions (IMFs) produced by EMD, and the basic principles of de-noising in the EMD domain.

In this chapter, we discuss the basics of EMD and its suitability in removing musical noise. Next, an optimum gain function is derived which minimizes the mean squared error (MSE) between the clean and estimated speech IMF variance in short frames. The individual IMFs are assumed to be normally distributed and thus the short-time variance is assumed to follow a Chi-square distribution. The gain is further modified to incorporate the speech presence uncertainty. An adaptive noise variance estimation method is also proposed utilizing the energy-period relationship of an IMF as found in [31].

4.1 Basics of EMD

The EMD is an adaptive decomposition method that separates a given signal $x(t)$, into a series of oscillating components, termed as intrinsic mode functions (IMFs)[2]. The IMFs are special functions that has symmetric envelopes with respect to the local mean. As the term implies, EMD is a heuristic transformation having no a priori basis function.

4.1.1 Intrinsic mode functions

An intrinsic mode function is a special kind of function proposed by Huang *et. al.* [2]. A function must satisfy two properties to be qualified as an IMF. It must (1) have the same numbers of zero crossings and extrema, and (2) has to be symmetric with respect to the local mean. These conditions restrict an IMF to have complex riding wave shapes and forces it to have a symmetric envelope. These properties are very useful when working on the Hilbert Transform of IMFs[2].

4.1.2 The sifting algorithm

The EMD process involves an iterative sifting algorithm that extracts the IMFs from the given signal. At first, the extremas of the given signal is found and curves are fitted through the maximas and the minimas. These curves are termed as the upper and lower envelopes, respectively. The mean value of these two envelopes is termed as the mean envelope. Since the objective of EMD is to make the mean envelope a constant zero (for satisfying the second IMF property), the mean envelope found in the first iteration is subtracted from the original signal. This is the basic idea of the EMD sifting method. After the subtraction, the original

signal will have lower fluctuations and more symmetry. However, rarely the IMF is found after the first subtraction. Thus, the process is continued iteratively. The new signal obtained after subtracting the first mean envelope is considered as the original signal and is processed again. After a number of iterations an IMF is obtained. The remaining IMFs can be obtained by repeating the process on the residual signal. The process is difficult to describe but can be easily understood from the algorithm steps given below.

Symbols

$x(t)$	Original signal to be decomposed
ϵ	A very small number used to set the stopping criteria
$\text{IMF}_j(t)$	The j th IMF
$r_j(t)$	The j th residual signal
$h_{j,i}(t)$	The j th approximation in the i th iteration
$U_{j,i}(t)$	The upper envelope of $h_{j,i}$
$L_{j,i}(t)$	The lower envelope of $h_{j,i}$
$\mu_{j,i}(t)$	The mean envelope of $h_{j,i}$
T	The total time duration.

Algorithm steps

Step-I Fix ϵ , $j \leftarrow 1$ (j th IMF)

Step-II $r_{j-1}(t) \leftarrow x(t)$ (residual)

Step-III Extract the j th IMF:

- (a) $h_{j,i-1}(t) \leftarrow r_{j-1}(t)$, $i \leftarrow 1$ (i number of sifts)
- (b) Extract local maxima/minima of $h_{j,i-1}(t)$
- (c) Compute the upper envelope and lower envelope functions $U_{j,i-1}(t)$ and $L_{j,i-1}(t)$ by interpolating respectively local maxima and minima of $h_{j,i-1}(t)$
- (d) Compute the mean of the envelope:

$$\mu_{j,i-1}(t) \leftarrow \frac{(U_{j,i-1}(t) + L_{j,i-1}(t))}{2}$$

- (e) Update: $h_{j,i}(t) \leftarrow h_{j,i-1}(t) - \mu_{j,i-1}(t)$, $i \leftarrow i + 1$

(f) Calculate stopping criterion:

$$SD(i) = \sum_{t=0}^T \frac{|h_{j,i-1}(t) - h_{j,i}(t)|^2}{|h_{j,i-1}(t)|^2}$$

(g) Decision: Repeat Step (b)-(f) until $SD(i) < \epsilon$ and then put $IMF_j(t) \leftarrow h_{j,i}(t)$ (j th IMF)

Step-IV Update residual: $r_j(t) \leftarrow r_{j-1}(t) - IMF_j(t)$

Step-V Repeat Step 3 with $j \leftarrow j + 1$ until the number of extrema in $r_j(t) \leq 2$.

The sifting is repeated several times (i) in order to get h to be a true IMF that fulfills the requirements (1) and (2). The result of the sifting procedure is that $x(t)$ will be decomposed into $IMF_j(t), j = 1, \dots, N$ and residual $r_N(t)$:

$$x(t) = \sum_{j=1}^N IMF_j(t) + r_N(t)$$

To guarantee that the IMF components retain enough physical sense of both amplitude and frequency modulations, a stopping criteria has to be maintained. This is accomplished by limiting the size of the standard deviation SD computed from the two consecutive sifting results. Usually, SD is set between 0.2 to 0.3 [2].

4.2 Separation of Musical Noise using EMD

It is known that the musical noise in the enhanced speech is composed of sinusoidal components with random frequencies that appear and disappear in each short-time frame. The generation of musical noise is mainly due to strong fluctuations of the noisy speech spectrum, especially in the noise dominated regions. These regions generate randomly spaced spectral peaks after the application of spectral attenuation [12]. We have observed that, if EMD is performed on such an enhanced speech and the IMFs are played back as sound files, the musical noise is distinctively heard in the first few IMFs. Thus, we presume that this noise can be efficiently separated using EMD.

To justify this intuition, an arbitrary speech file taken from the TIMIT database is corrupted by white noise at 10dB SNR, and 512 point DCT is taken on each

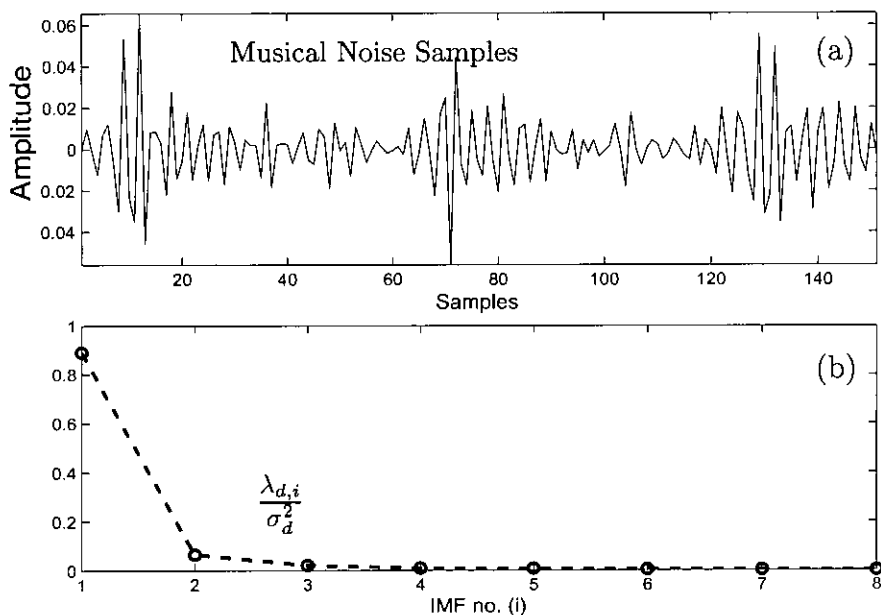


Fig. 4.1: (a) A sequence of musical noise. (b) The energy distribution of musical noise in different IMFs (ratio of the i th IMF variance to the overall signal variance).

32ms frame. In each frame, the higher 256 DCT coefficients (257 to 512) are enhanced employing the traditional Wiener filter in DCT domain with the optimum *a priori* SNR [27] and the lower 256 coefficients are replaced by the clean speech DCT coefficients. This enhanced speech is subtracted from the clean speech, to have only the high frequency residual noise. This is shown in Fig. 4.1(a). From this figure, it can be observed that the musical noise is very much similar to typical IMFs, having non-stationary oscillations [2]. We performed EMD on this musical noise and the variance of each IMF was calculated. Fig. 4.1(b) shows the ratio of the i th IMF variance $\lambda_{d,i}$ and the overall musical noise variance λ_d plotted against the IMF number. This depicts the fact that most of the musical noise energy (88.94%) is concentrated in the first IMF. This justifies our assumption that the musical noise can be well separated by EMD into IMFs. Hence, the musical artifacts may be effectively filtered in the EMD domain by attenuating the noise dominated regions.

4.3 Proposed method

If $y(n)$, $x(n)$ and $d(n)$ denote the noisy speech (output of the first stage), clean speech and the residual noise, respectively, we may write,

$$y(n) = x(n) + d(n). \quad (4.1)$$

Since EMD is a heuristic transformation, no direct relation between the signal and noise IMFs is available. Also, because the process is nonlinear,

$$y_i(n) \neq x_i(n) + d_i(n)$$

inequality holds, where $y_i(n)$, $x_i(n)$, and $d_i(n)$ denote the i th IMF of the noisy signal, clean signal and residual noise, respectively. This nonlinearity makes EMD domain analysis very difficult. Unless an analytic expression of EMD is found, deriving a suppression rule for processing time samples of the noisy IMFs one by one is not feasible. We, therefore, propose to process IMFs in non overlapping segments instead. Before developing the segment-wise filtering scheme, we attempt to relate the variances of signal and noise IMFs in short time segments.

4.3.1 Conservation of energy in EMD

As mentioned earlier, an analytic expression relating $x_i(n)$, $d_i(n)$ and $y_i(n)$ is not available till date. However, an approximate relation between their short time variance can be found experimentally, which is given by,

$$\lambda_y(i, k) \approx \lambda_x(i, k) + \lambda_d(i, k) \quad (4.2)$$

where, $\lambda_y(i, k)$, $\lambda_x(i, k)$ and $\lambda_d(i, k)$ are the variances of the i th IMF of noisy speech, clean speech, and residual noise in the k th short-time segment. This approximate relation is used in our method which greatly simplifies the derivation of the proposed suppression rule. Yet, we have numerically investigated this conjecture. A speech utterance from the TIMIT database is corrupted by white noise so as to give 0, 10 and 20dB SNR, and then enhanced by the Wiener filter in the DCT domain using the variable averaging parameter proposed in [27]. For each enhanced speech, clean speech and the residual noise, EMD is computed upto 4 IMFs to find $x_i(n)$, $y_i(n)$ and $d_i(n)$. All the IMFs are segmented in sizes from 64 to 1024 samples with an increment of 64 samples. In all of these

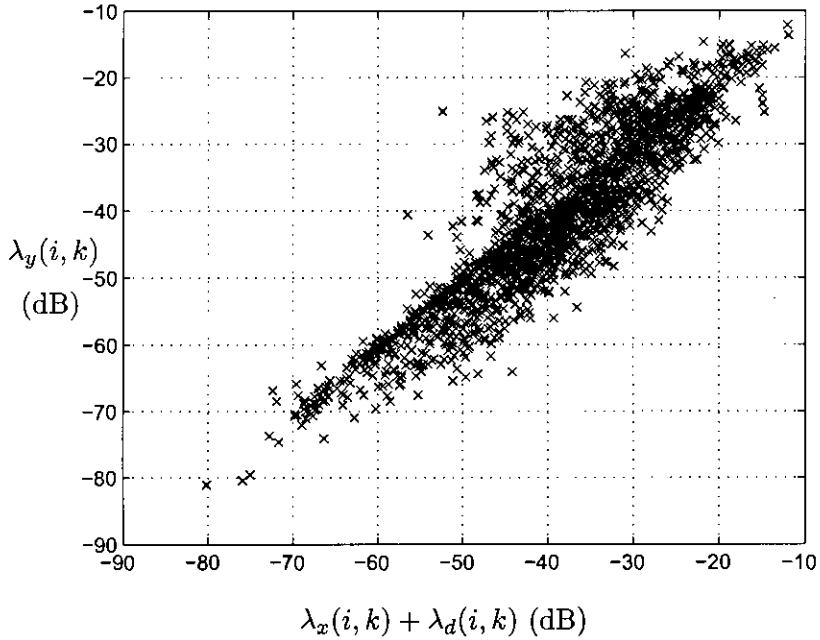


Fig. 4.2: Scatter plot of $\lambda_y(i, k)$ vs. $\lambda_x(i, k) + \lambda_d(i, k)$ in Log scale.

frames, $\lambda_x(i, k)$, $\lambda_d(i, k)$ and $\lambda_y(i, k)$ are determined to generate the scatter plot of Fig. 4.2. The clearly visible linear trend inspired us to use the approximation of (4.2).

4.3.2 Statistical model of IMF local variance

The energy, $z = \sum_{n=1}^N u^2(n)$, of a normally distributed random variable $u(n) \sim \mathcal{N}(0,1)$ in a short segment of length N samples follows a Chi-square distribution having N degrees of freedom [26]. The probability density function (PDF) of z is given by,

$$p(z; N) = \frac{1}{2^{N/2}\Gamma(\frac{N}{2})} z^{\frac{N}{2}-1} e^{-z/2}. \quad (4.3)$$

Thus, if an arbitrary IMF of a Gaussian sequence has an overall variance $\bar{\lambda}$ and a local variance λ , in a short-time region of length N , the normalized IMF energy $N\lambda/\bar{\lambda}$ will follow a Chi-square density function given by (4.3), as discussed in [31]. Thus, the PDF of λ is found to be,

$$p(\lambda) = \frac{N(N\lambda/\bar{\lambda})^{\frac{N}{2}-1}}{2^{N/2}\Gamma(\frac{N}{2})\bar{\lambda}} \exp\left(-\frac{N\lambda}{2\bar{\lambda}}\right). \quad (4.4)$$

Assuming that $y(n)$, $x(n)$ and $d(n)$ are Gaussian [32], their IMFs can also be assumed to be Gaussian [31] and thus the PDF of (4.4) may be used for the IMF

local variances $\lambda_y(i, k)$, $\lambda_x(i, k)$ and $\lambda_d(i, k)$ defined in Section 4.3.1.

4.3.3 Optimum gain

To formulate a suppression rule in the energy domain, we aim to minimize the cost function

$$J_G = E\{(\hat{\lambda}_x(i, k) - \lambda_x(i, k))^2\}, \quad (4.5)$$

where, $\hat{\lambda}_x(i, k) = G_\lambda(i, k)\lambda_y(i, k)$ is the estimate of clean IMF local variance. The indices i and k denote the IMF and frame number, respectively. Using the approximation in (4.2), the gain $G_\lambda(i, k)$ minimizing J_G is given by

$$G_\lambda = \frac{E\{\lambda_x^2\} + E\{\lambda_x\lambda_d\}}{E\{\lambda_x^2\} + E\{\lambda_d^2\} + 2E\{\lambda_x\lambda_d\}}. \quad (4.6)$$

The indices i and k are omitted for simplicity. Assuming that λ_x and λ_d are independent and they follow a Chi-square density function as given in (4.4), we may substitute,

$$E\{\lambda_x^2\} = \left(1 + \frac{2}{N}\right) \bar{\lambda}_x^2 \quad (4.7)$$

$$E\{\lambda_d^2\} = \left(1 + \frac{2}{N}\right) \bar{\lambda}_d^2 \quad (4.8)$$

and $E\{\lambda_x\lambda_d\} = \bar{\lambda}_x\bar{\lambda}_d$.

Thus, using these definitions, (4.6) gives the gain expression as

$$G_\lambda(i, k) = \frac{\bar{\xi}_{i,k}^2 + \frac{\bar{\xi}_{i,k}}{2/N_{i,k}+1}}{\bar{\xi}_{i,k}^2 + 1 + \frac{2\bar{\xi}_{i,k}}{2/N_{i,k}+1}}, \quad (4.9)$$

where,

$$\bar{\xi}_{i,k} = \text{SNR}_{a \text{ priori}} = \frac{\bar{\lambda}_x(i, k)}{\lambda_d(i, k)}. \quad (4.10)$$

We propose to segment the IMFs at the zero-crossing points containing integer number of oscillation periods. Thus the frame size $N_{i,k}$ is different for each frame.

The gain in (4.9) gives a variance estimate. Therefore, a square root operation must be performed in order to apply it on short time segments. The final gain expression is thus,

$$G(i, k) = \sqrt{\frac{\bar{\xi}_{i,k}^2 + \frac{\bar{\xi}_{i,k}}{2/N_{i,k}+1}}{\bar{\xi}_{i,k}^2 + 1 + \frac{2\bar{\xi}_{i,k}}{2/N_{i,k}+1}}}. \quad (4.11)$$

The *a priori* SNR in (4.11) can be calculated using the classical decision directed approach [9] applied in time-domain:

$$\bar{\xi}_{i,k} = \beta \bar{\xi}_{i,k-1} + (1 - \beta) \max[0, \gamma_{i,k} - 1], \quad (4.12)$$

where β is a smoothing parameter and,

$$\gamma_{i,k} = \text{SNR}_{post} = \frac{\lambda_y(i, k)}{\bar{\lambda}_d(i, k)}.$$

For the first frame, $\bar{\xi}_{i,1} = \max[0, \gamma_{i,1} - 1]$ is used.

4.3.4 Considering speech presence uncertainty

If the events H_1 and H_0 denote speech presence and absence, respectively, from (4.4) we have,

$$p(\lambda_y|H_1) = \frac{N(N\lambda_y/\bar{\lambda}_y)^{\frac{N}{2}-1}}{2^{N/2}\Gamma(\frac{N}{2})\bar{\lambda}_y} \exp\left(-\frac{N\lambda_y}{2\bar{\lambda}_y}\right), \quad (4.13)$$

$$p(\lambda_y|H_0) = \frac{N(N\lambda_y/\bar{\lambda}_d)^{\frac{N}{2}-1}}{2^{N/2}\Gamma(\frac{N}{2})\bar{\lambda}_d} \exp\left(-\frac{N\lambda_y}{2\bar{\lambda}_d}\right). \quad (4.14)$$

Using Bayes rule, the *a posteriori* probability for speech presence is given by,

$$p(H_1|\lambda_y) = \frac{1}{1 + \Lambda} \triangleq G_{pr},$$

where,

$$\Lambda \triangleq \frac{p(H_0)p(\lambda_y|H_0)}{p(H_1)p(\lambda_y|H_1)}. \quad (4.15)$$

Assuming that the speech and noise states are equally likely, i.e., $p(H_1) = p(H_0) = \frac{1}{2}$, from (4.13), (4.14) and (4.15) we have,

$$\Lambda = (1 + \bar{\xi})^{N/2} \exp\left(-\frac{N}{2} \frac{\gamma \bar{\xi}}{(\bar{\xi} + 1)}\right). \quad (4.16)$$

Thus, the modified gain expression is given by

$$G_{opt}(i, k) = G_{pr}G(i, k). \quad (4.17)$$

Incorporating G_{pr} , the gain is now highly attenuating for low $\bar{\xi}_{i,k}$ values. This property is highly effective for musical noise reduction.

4.3.5 Noise variance estimation

To determine the noise variance in different IMFs, EMD is applied in the speech absence periods of the noisy speech. From these frames, the average noise variance $\bar{\lambda}_{d,i}$ and mean period $\bar{T}_{d,i}$ of the noise IMFs [31] are obtained. It is shown in [31] that, for a Gaussian sequence, IMF energy per period is a constant, i.e.,

$$\bar{\lambda}_{d,i}\bar{T}_{d,i} = K, \quad (4.18)$$

where K is a constant. Since EMD does not allow same time scale data to be present at the same location in different IMFs [2], intuitively, noise embedded in a region of a noisy IMF cannot have a drastically different local time scale. Thus, we assume $\bar{T}_d(i, k) \approx \bar{T}_y(i, k)$, where $\bar{T}_d(i, k)$ and $\bar{T}_y(i, k)$ denote the average period of $d_i(n)$ and $y_i(n)$ in the k th frame, respectively. Thus the noise variance may be estimated as

$$\hat{\lambda}_d(i, k) = \frac{K}{\bar{T}_y(i, k)}. \quad (4.19)$$

In reality, the right hand side of (4.18) may give slightly different value for different IMFs. Thus, we calculate K as

$$K = \frac{1}{M-1} \sum_{i=2}^M \bar{\lambda}_{d,i}\bar{T}_{d,i}, \quad (4.20)$$

where M is the total number of IMFs. Since, (4.18) is not valid for the first IMF [31], the summation begins at $i = 2$. Thus, $\bar{\lambda}_d(1, k)$ is calculated from the speech absence periods directly.

4.4 Simulation results

4.4.1 Experimental details

The effectiveness of the proposed post processing scheme is evaluated using 5 male and 5 female utterances taken from the TIMIT database. Two different types of noise, e.g. ‘white’ and ‘babble’ were taken from the NOISEX database to corrupt the speech signals. The sampling frequency was 16 kHz. The noisy signals were enhanced using the i) Wiener filter in DCT-domain using the adaptive α [27], ii) approximate MAP [10] with $\alpha = 0.98$, and iii) approximate MAP [10] with the adaptive α [27], iv) the MMSE STSA estimator [9], v) the MMSE

Log STSA estimator [8], and vi) the perceptual filtering proposed by Virag [16]. These methods are denoted by: i) Wn (α), ii) MAP (0.98), iii) MAP (α), iv) EM STSA, iv) EM Log STSA and v) Virag, respectively. The proposed post processor is applied after the Wn (α) and MAP (α). After processing, they are denoted by P-Wn (α) and P-MAP (α), respectively.

The proposed post processor was applied in the following procedure. EMD was applied on the enhanced speech to extract the IMFs and the lowest 4 IMFs were taken. The zero-crossing points of these IMFs were identified, and segmentation was done using two full oscillation cycles. No overlapping was used. For each segment, the value of $\xi_{i,k}$ was calculated from (4.12) using $\beta = 0.95$. The average noise variance $\bar{\lambda}_{d,i}$ and the noise variance in each frame $\hat{\lambda}_d(i, k)$ were found as described in Section 4.3.5. The non-speech regions were detected using the method in [33]. Next, the gain $G_{opt}(i, k)$ was calculated from (4.17) and applied to the short frames. Simple summation of all the modified (and unmodified) IMFs gives the further enhanced signal.

Comparison of the objective quality measures for a wide range of input SNRs is shown in Figs. 4.3, 4.4, 4.4.1. In Fig. 4.3, we have compared the performance of MAP(0.98), MAP(α), P-MAP(α) and EM STSA. The average segmental SNR (AvgSegSNR) and composite speech quality measure (COMP) [34] were used for objective performance comparison. In Figs. 4.4, 4.4.1 we have compared the performance of MAP(0.98), MAP(α), P-MAP(α), Wn(α), P-Wn(α), EM Log STSA and EM STSA with respect to the quality indices: Overall SNR, AvgSegSNR, PESQ and COMP.

The COMP quality index is a linear combination of different objective quality measures. The correlation coefficients for the linear combination are determined from listening tests [34]. The index is given by,

$$\text{COMP} = 1.594 + 0.805\text{PESQ} - 0.512\text{LLR} - 0.007\text{WSS}.$$

This is known to be highly correlated with human listening.

For subjective quality evaluation, two listening tests were performed. The first one was conducted using the comparison category rating (CCR) method [35]. A total number of 10 listeners were presented with the enhanced speech files in pairs and asked to compare their quality in the CCR scale [35]. The order of presentation was random. To eliminate any biasing due to the order of the

algorithms within a pair, each pair of enhanced utterances was presented twice, with the order switched. Listeners were asked to rate the quality of the second utterance relative to that of the first according to the scale in Table 4.4.1.

For the second listening test, the MOS (Mean Opinion Score) scale was used. This scale is given in Table 4.4.1. 2 male and 2 female utterances from the TIMIT database corrupted by ‘white’ and ‘babble’ noise at 5 and 10dB SNR was enhanced by the Wiener DCT [27] and the EM-Log STSA [8] method and the proposed EMD based post-filter was applied. A total of 8 subjects attended the informal listening test and ranked the speech quality in the MOS scale. The averaged results are given in Table 4.4.

Table 4.1: The comparison category rating (CCR) scale

3	Much better
2	Better
1	Slightly better
0	About the same
-1	Slightly worse
-2	Worse
-3	Much worse

Table 4.2: The mean opinion score (MOS) scale

5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

4.4.2 Performance Evaluation and Discussion

From Figs. 4.3 (a) and (c), it is clear that MAP (α) is superior compared to MAP (0.98) [10] with respect to the AvgSegSNR. This is true throughout the whole input SNR range and for both noise types. However, this improvement is accompanied by musical noise generation. After the EMD based post processing, further improvement in AvgSegSNR and COMP is achieved for a wide input SNR range. This can be observed from Figs. 4.3 (a) and (c), and, (b) and (d), respectively. Interestingly, this improvement is accompanied by a simultaneous improvement in listening quality, which can be seen in the fourth row of Table 4.3. The positive scores indicate the preference of P-MAP (α) over MAP (α). Furthermore, the P-MAP (α) and MAP (0.98) are very close in terms of the listening quality, as can be seen from the fifth row of Table 4.3. As evident from these results, the proposed technique can be used for improving objective quality without

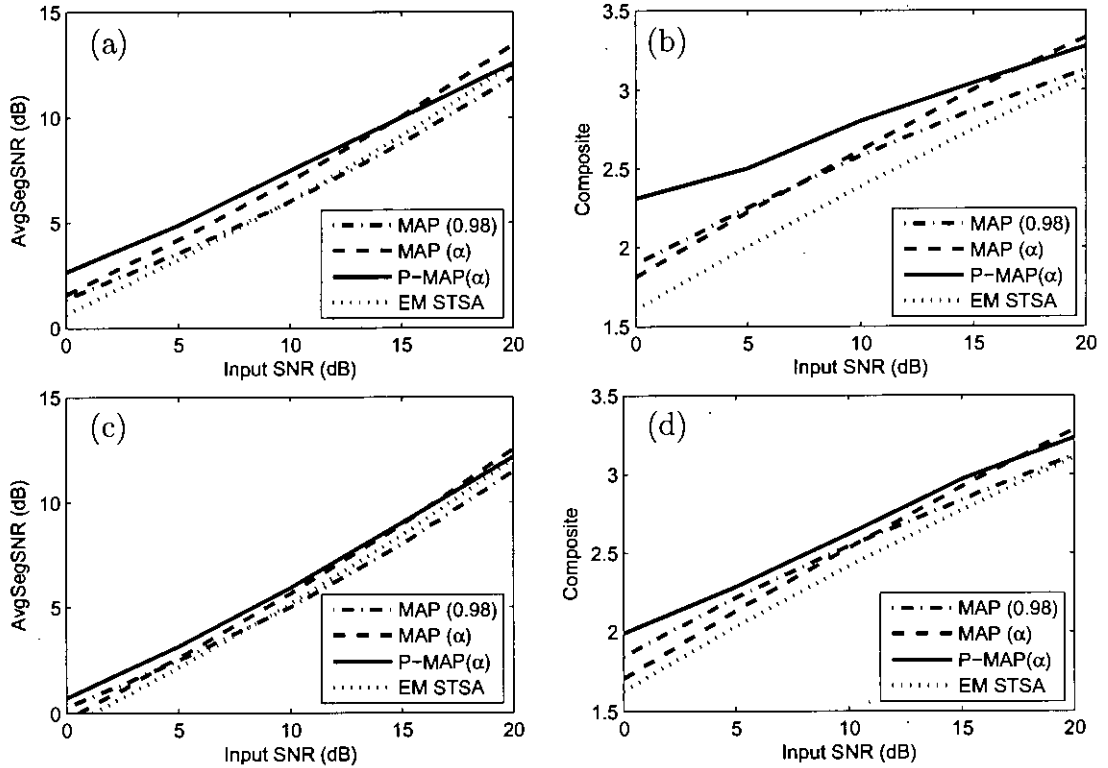


Fig. 4.3: Average objective quality measures with different input SNRs; (a), (b): white noise; (c), (d): babble noise.

deteriorating the subjective quality of the enhanced speech. The improvement in listening quality after the post filtering is also demonstrated from the MOS scores given in Table 4.4.

A more detailed comparison of the proposed post filtering method is presented in Figs. 4.4 and 4.4.1. From these figures, the superiority of the proposed post filtering technique with respect to AvgSegSNR is also apparent. However, the post filtering operation reduces the overall SNR for input SNR greater than 4dB. Nevertheless, the remarkable improvement in the PESQ and COMP scores demonstrate the superior listening quality of the proposed enhancement scheme.

It may be emphasized that the proposed method improves the quality indices, that is the CCR scores and the MOS scores, without sacrificing the AvgSegSNR values. This is a very important finding since it is contrary to general observations of speech enhancement techniques. Moreover, the proposed method is providing a similar listening quality in the enhanced speech that has a superior objective quality. Thus, it may be concluded that the proposed method maybe leading the speech enhancement research to a new level of speech quality.

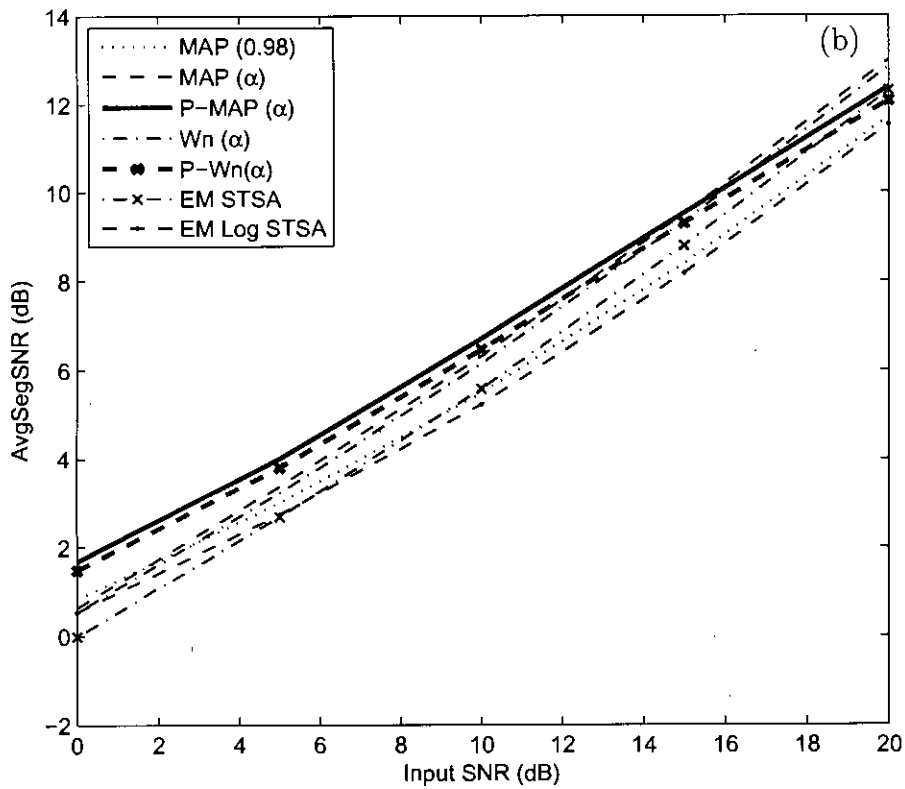
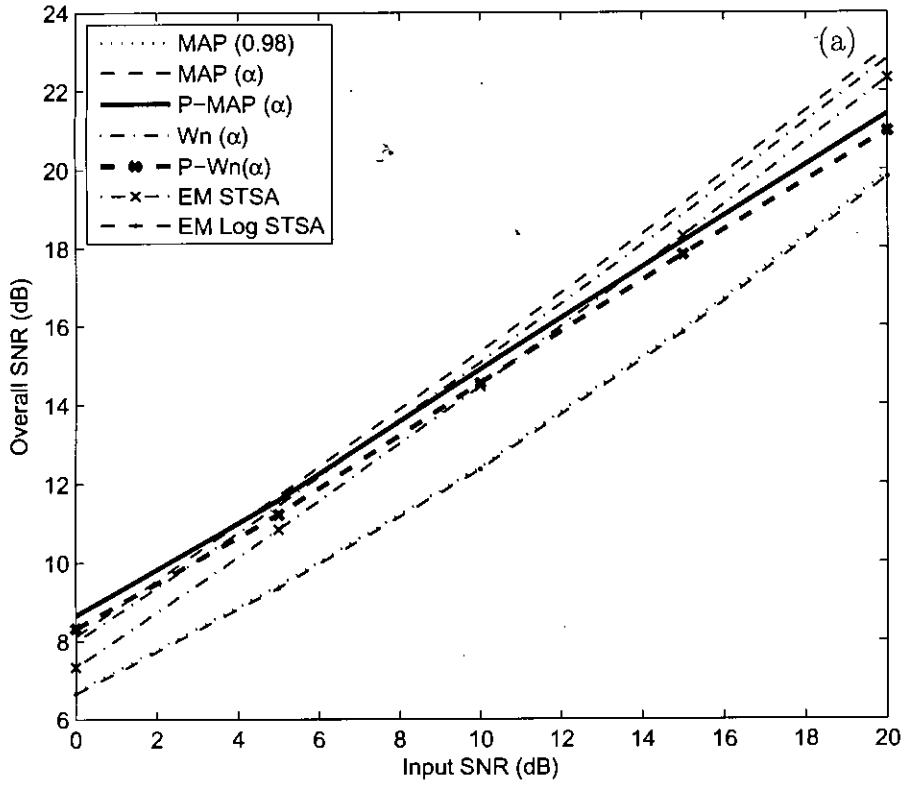


Fig. 4.4: Average objective quality measures with different input SNRs. Results obtained from 'white' and 'babble' noise are averaged.

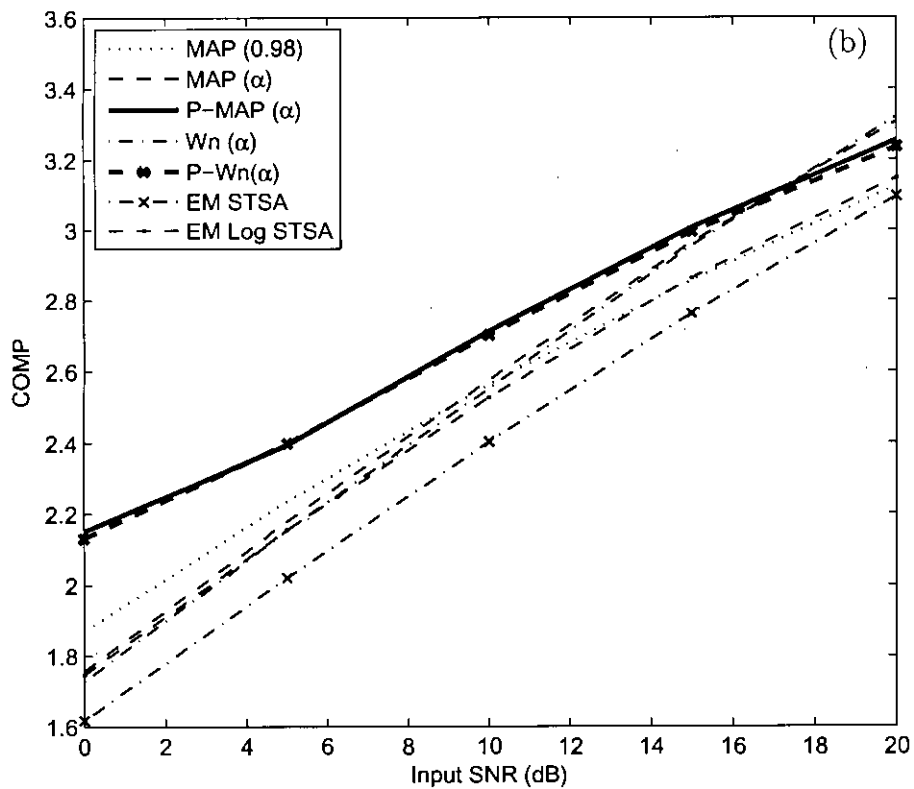
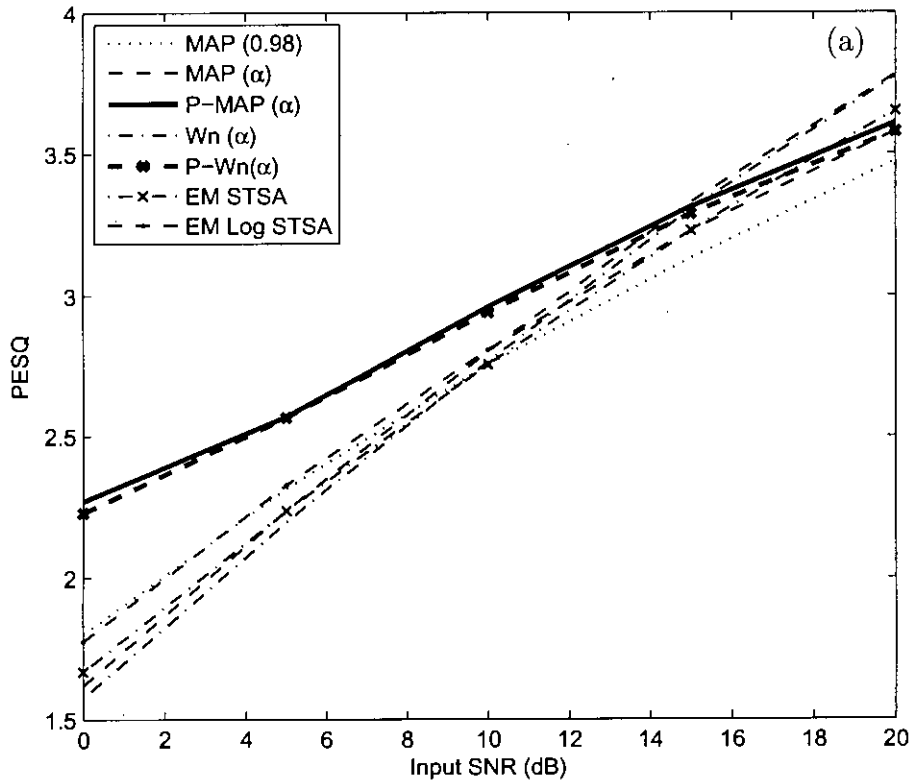


Fig. 4.5: Average objective quality measures with different input SNRs. Results obtained from 'white' and 'babble' noise are averaged.

Table 4.3: Results from the Listening test in the CCR scale. A positive value indicates preference for the proposed method

Noise Type	white		babble	
	5dB	10dB	5dB	10dB
P-MAP (α) vs MAP (α)	1.5000	1.7500	1.2000	1.3250
P-MAP (α) vs MAP (0.98)	0.9500	0.8750	0.8250	0.8000
P-Wn (α) vs Wn (α)	1.4250	1.7000	0.4750	1.3250
P-Wn (α) vs Virag	1.4500	1.7250	1.2750	1.0750

Table 4.4: The Mean Opinion Score Experimental Results for white and babble noise of 5 and 10dB SNR

Noise Type	white		babble		
	5dB	10dB	5dB	10dB	
Noisy Signal	1.16	1.31	1.16	1.34	
EM Log-STSA [8]	Single Stage	2.38	2.88	2.06	2.63
	Proposed	3.09	3.56	2.44	3.28
Wiener DCT [27]	Single Stage	2.34	2.81	2.31	2.66
	Proposed	3.00	3.84	2.44	3.31

4.5 Conclusion

This chapter has dealt with a novel post processing method using EMD for suppression of residual noise from speech signals. Well known techniques in the DCT and DFT domain were used as the first stage filter and the annoying musical noise present in the enhanced speech were filtered in the second stage using the proposed EMD domain method. Experimental results have demonstrated the superiority of our method in terms of both objective and subjective evaluation criteria.

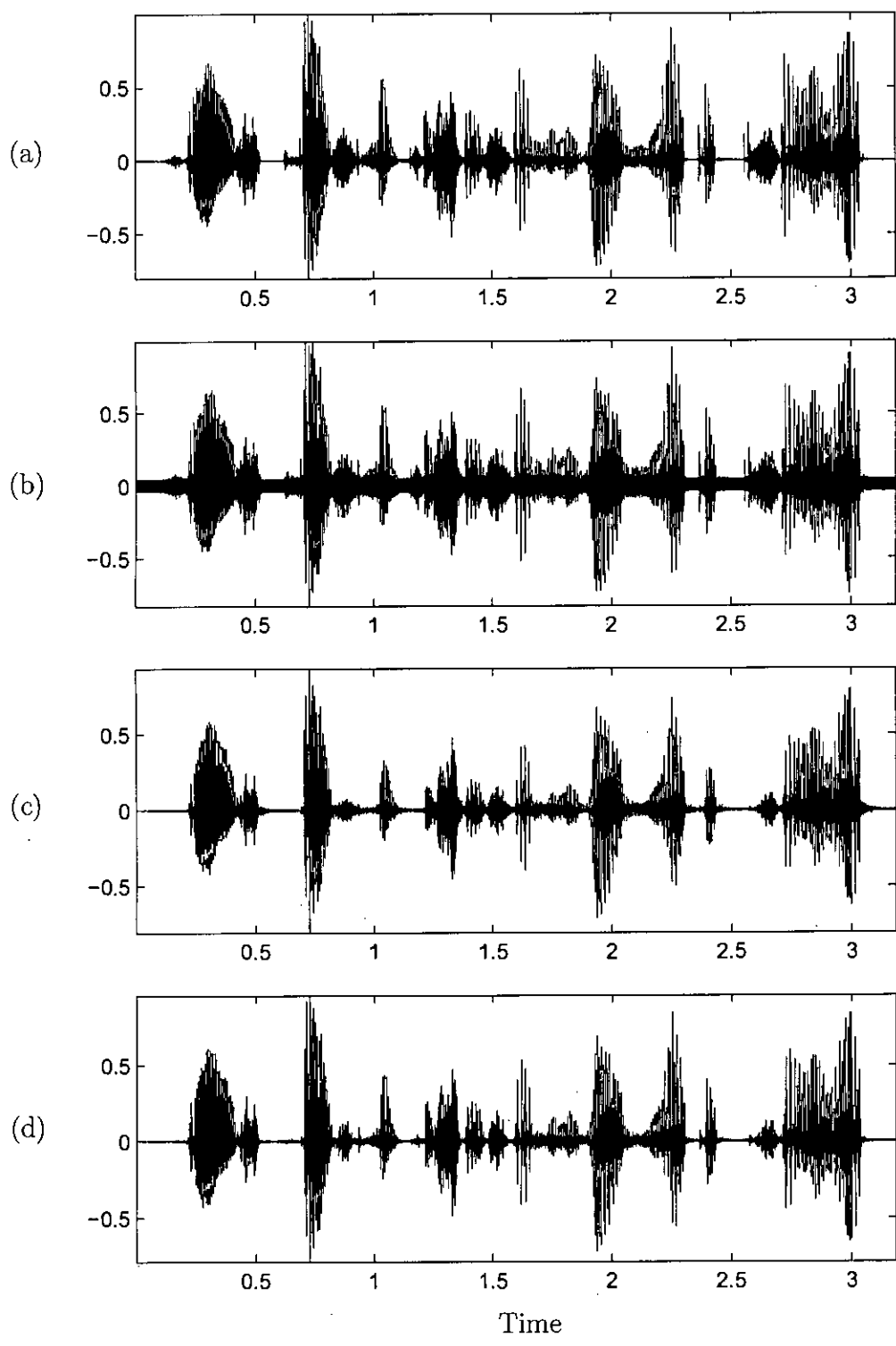


Fig. 4.8: Enhancement results for the male utterance “Heels place emphasis on the long legged silhouette”. Time domain plots of (a) clean speech, (b) noisy speech (10dB), (c) enhanced using MAP(α) and (d) enhanced using P-MAP(α).

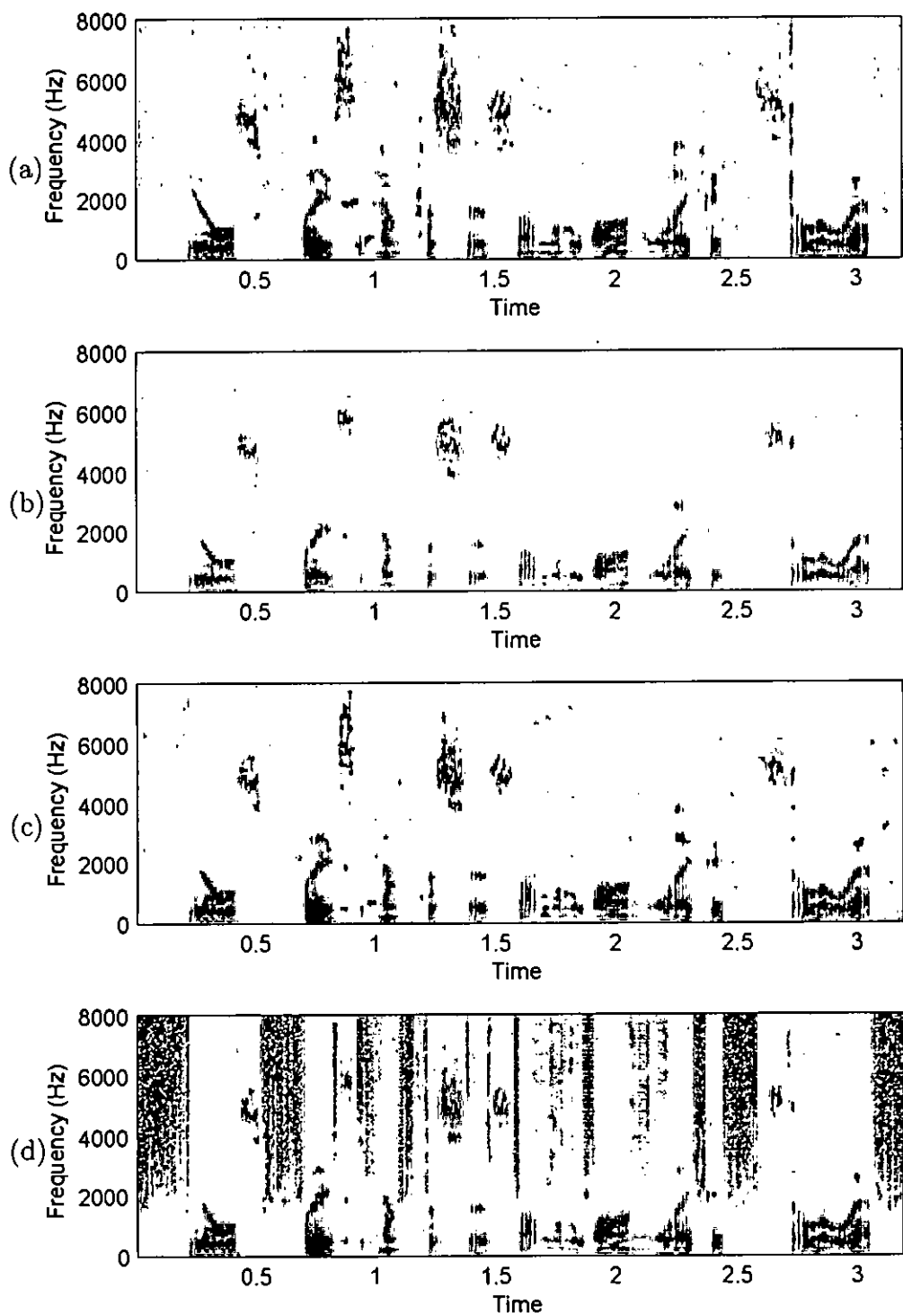


Fig. 4.7: Enhancement results for the male utterance “Heels place emphasis on the long legged silhouette”. Spectrogram plots of (a) clean speech, (b) noisy speech (10dB), (c) enhanced using $W_n(\alpha)$ and (d) enhanced using $P-W_n(\alpha)$.

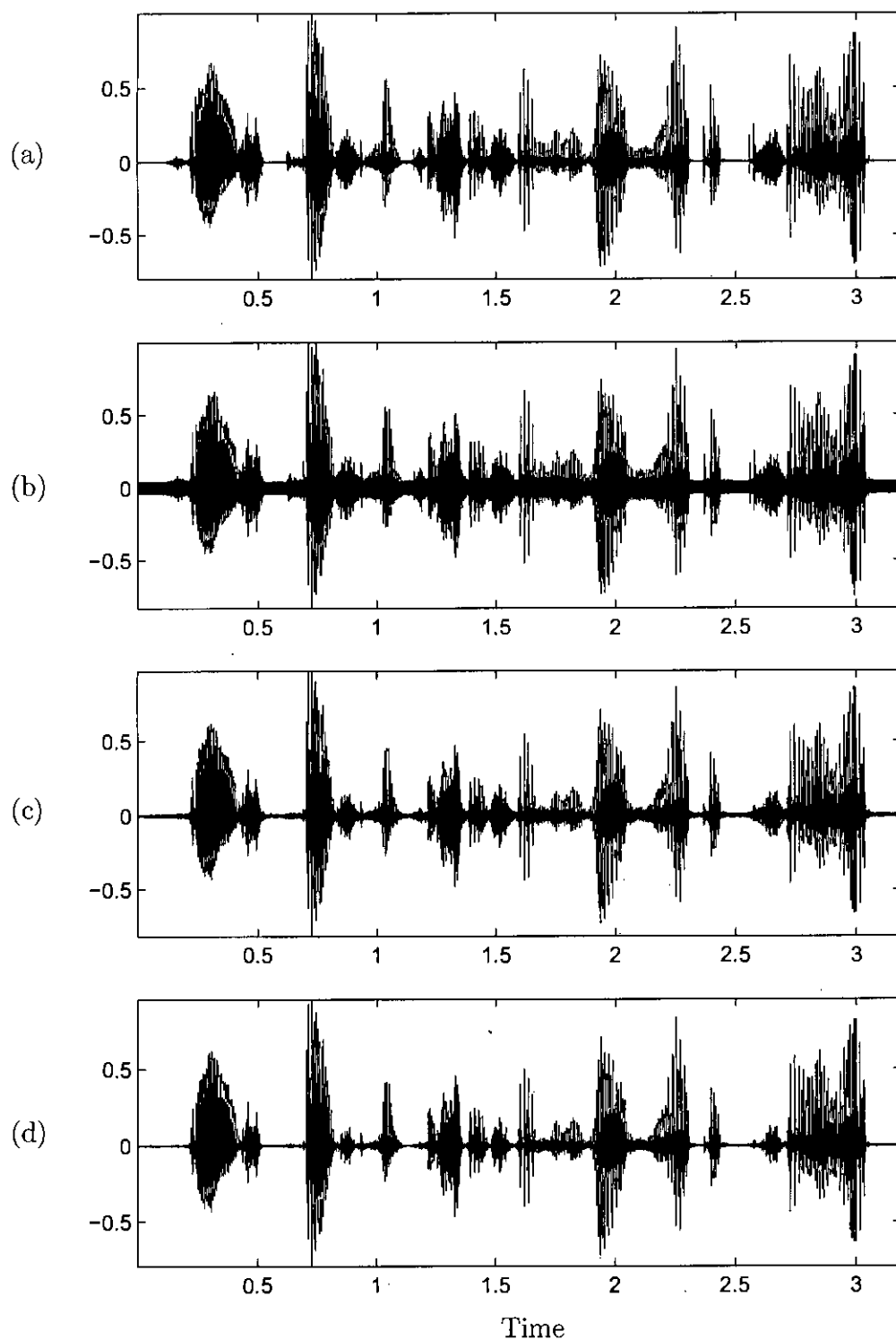


Fig. 4.6: Enhancement results for the male utterance “Heels place emphasis on the long legged silhouette”. Time domain plots of (a) clean speech, (b) noisy speech (10dB), (c) enhanced using $W_n(\alpha)$ and (d) enhanced using $P-W_n(\alpha)$.

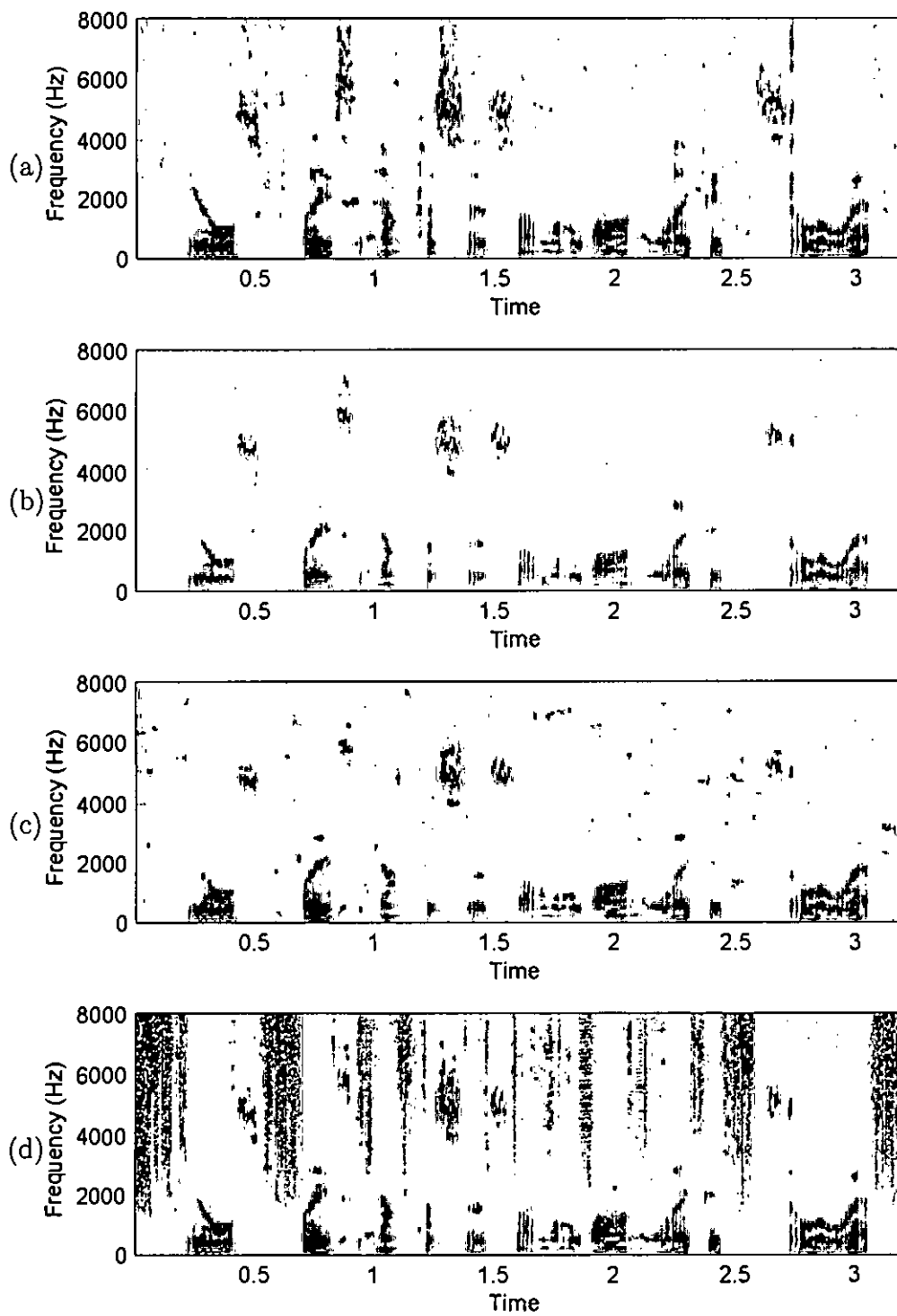


Fig. 4.9: Enhancement results for the male utterance “Heels place emphasis on the long legged silhouette”. Spectrogram plots of (a) clean speech, (b) noisy speech (10dB), (c) enhanced using $MAP(\alpha)$ and (d) enhanced using $P-MAP(\alpha)$.

Chapter 5

Performance analysis of the hybrid method

5.1 Introduction

In this chapter, the performance of the proposed hybrid speech enhancement algorithm is studied and compared to that of the dual gain Wiener (DGW) [3] filter, the conventional Wiener filter and the MMSE Log spectral amplitude estimator [8]. The quality of the enhanced speech is measured using the overall SNR, average segmental SNR and PESQ (perceptual evaluation of speech quality) [36].

5.2 Experimental details

The effectiveness of the proposed hybrid enhancement scheme is evaluated using 10 male and 10 female utterances taken from the TIMIT database. The speech files are corrupted by white noise taken from the NOISEX database. The noise level is adjusted so as to give SNR from 0 dB to 25 dB. The sampling frequency is 16kHz. A frame size of 32ms (512 samples) is used for framing and the overlap-add method with 75% overlap is used for signal decomposition. The *a priori* SNR is calculated using decision directed approach [9] utilizing the variable averaging parameter proposed in [27].

In the first stage, the noisy speech is enhanced using the conventional Wiener (in DCT), dual gain Wiener (DGW), MMSE Log spectral amplitude estimator (EM Log STSA) [8], and the proposed dual MMSE (DMMSE) estimator. Each frame is subjected to 1024 point DCT for the Wiener, DGW, and the DMMSE estimator. The experimental methods of the DMMSE implementation is discussed

in Section 3.4.2 of this thesis. In the second stage, the proposed EMD based post residual noise reduction technique was applied to the speech files enhanced using the conventional Wiener, DMMSE and EM Log STSA. The experimental details regarding the post processing was done as discussed in Section 4.4. The average of the results of the 20 utterances are plotted in Figs. 5.1, 5.2 and 5.3.

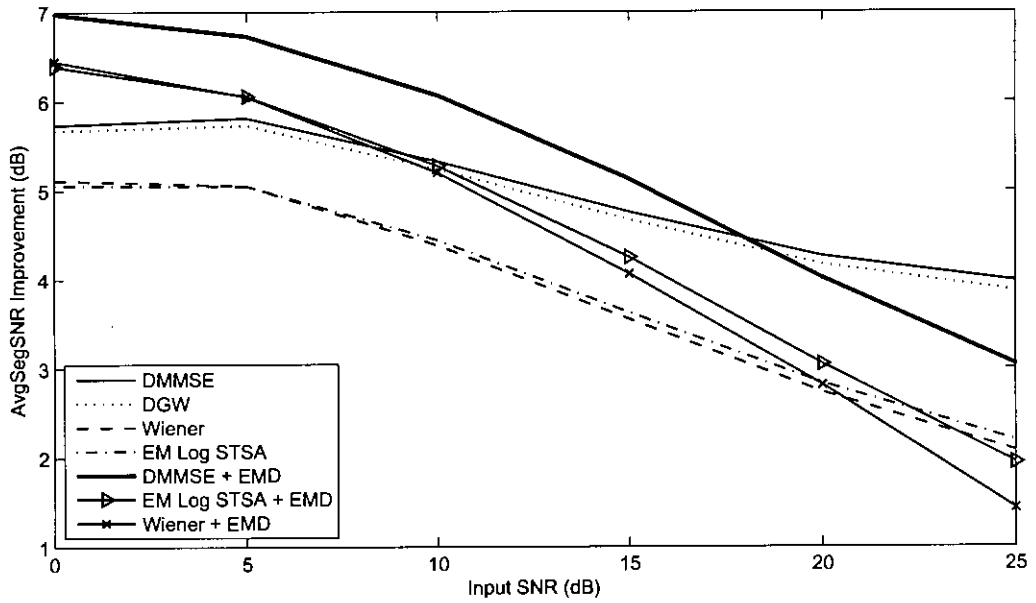
5.3 Performance comparison and discussion

As mentioned in section 4.4, the average segmental SNR (AvgSegSNR) improvement has been the most remarkable for the proposed post filtering algorithm. This is clearly demonstrated in Figs. 5.1 (a) and (b), especially in the SNR regions below 15 dB for all the methods. The post processing method fails to manifest further improvement above 15dB, for DMMSE and above 20dB for Wiener and EM Log STSA as one can see from Fig. 5.1. For a polarity estimator of accuracy 80%, the situation is almost similar, except that the DMMSE and DGW performances are reduced, which is expected. However, the relative superiority of DMMSE compared to DGW is still visible.

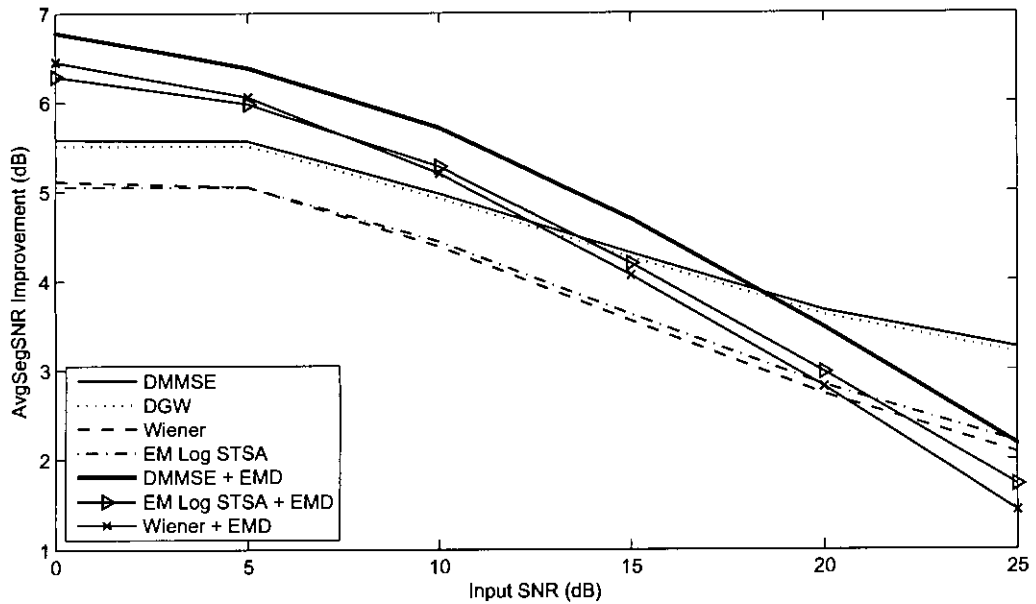
Figs. 5.2 (a) and (b) show the overall performance of the various methods interms of the output SNR improvement. Here, while the DMMSE, and DGW demonstrate superior performance than the Wiener and EM Log STSA estimators in most of the regions, the application of the proposed post filtering method reduces the improvement for input SNRs greater than 10dB. Similar behavior is observed in case of Figs. 5.2 (a) and (b). However, the PESQ improvement demonstrated in Figs. 5.2 (a) and (b) show noticeable perceptual quality improvement of the enhanced speech files. Even though the proposed post filtering technique reduces the overall SNR in some regions, the listening quality improvement obtained using the method is considerable. This is expected, since it is a well known fact that overall SNR is not correlated to human listening [34]. This was also confirmed earlier in the MOS experiments discussed in section 4.4. The PESQ scores in this case gives us an appraisal of that improvement. In this case also, the performance of the post filtering method deteriorates at very high SNRs (> 20dB). But the improvement in the PESQ score using the hybrid method (DMMSE+EMD) is remarkable in the lower SNR regions. This is true in both the ideal and simulated (polarity estimation accuracy of 80%) cases.

5.4 Conclusion

In this chapter, the performance of the hybrid algorithm has been analyzed and compared to well established speech enhancement techniques. The proposed dual MMSE estimator shows superior objective quality indices compared to other methods in most of the input SNR regions. Significant improvement in the average segmental SNR values has been achieved after the application of the proposed post-processing technique.

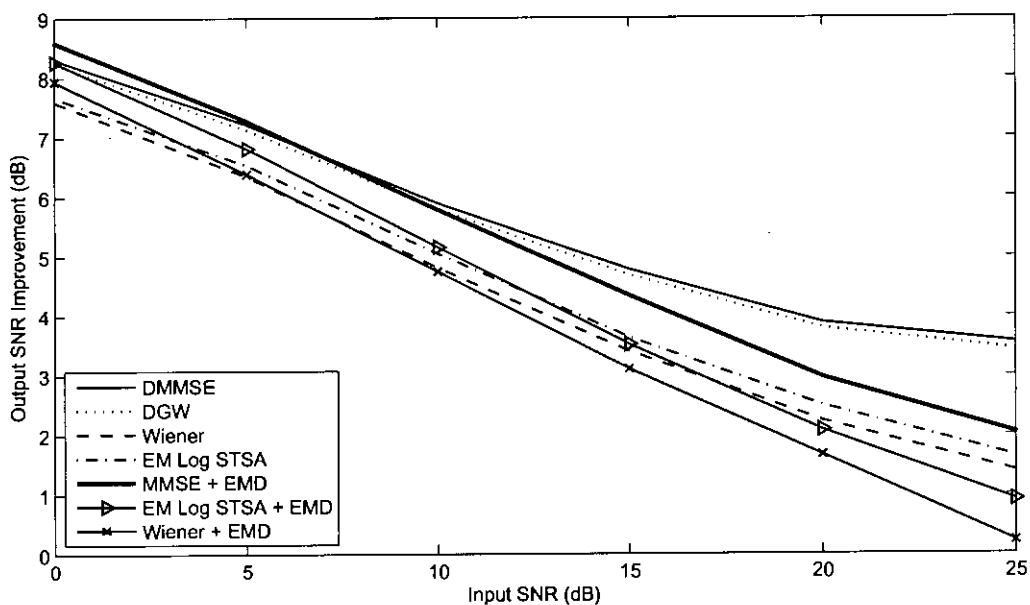


(a) Results using a polarity estimator of 100% accuracy (The ideal case).

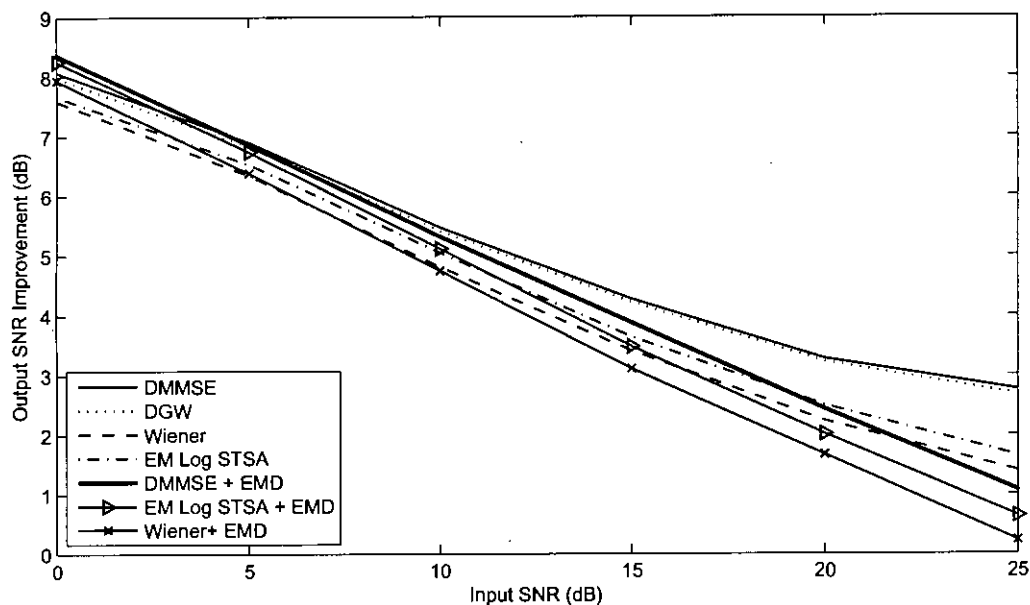


(b) Results using a polarity estimator of 80% accuracy.

Fig. 5.1: Performance comparison of the DMMSE, DGW, Wiener and EM Log STSA with respect to improvement in average Segmental SNR for an input SNR range of 0 dB to 25 dB. The proposed post filtering method is applied to the DMMSE, Wiener and EM Log STSA methods indicated by the +EMD notation. DMMSE+EMD indicates the proposed hybrid method. A polarity estimator of accuracy 100% and 80% was used in (a) and (b), respectively.

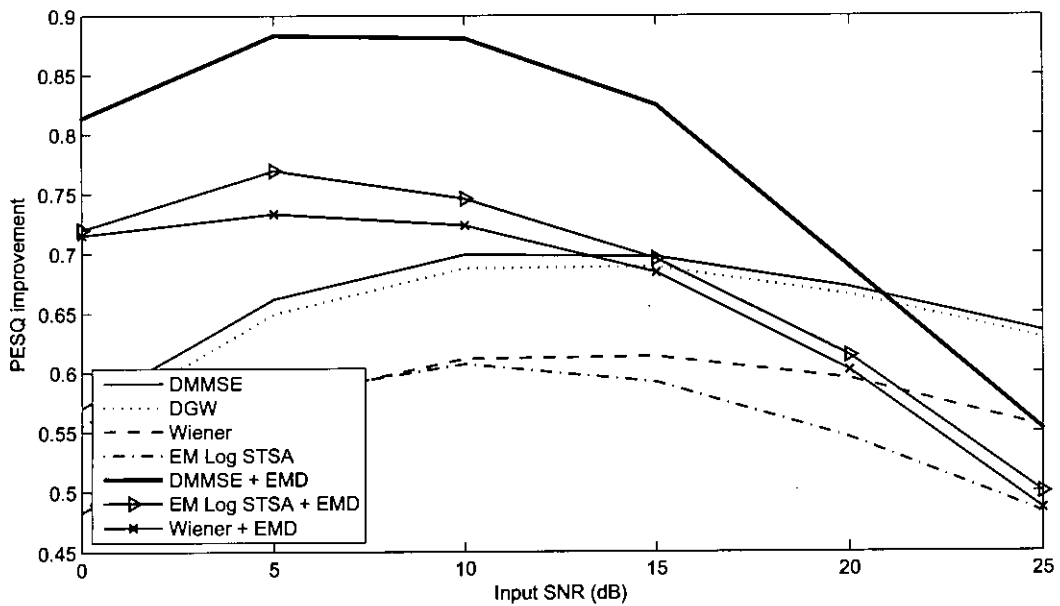


(a) Results using a polarity estimator of 100% accuracy (The ideal case).

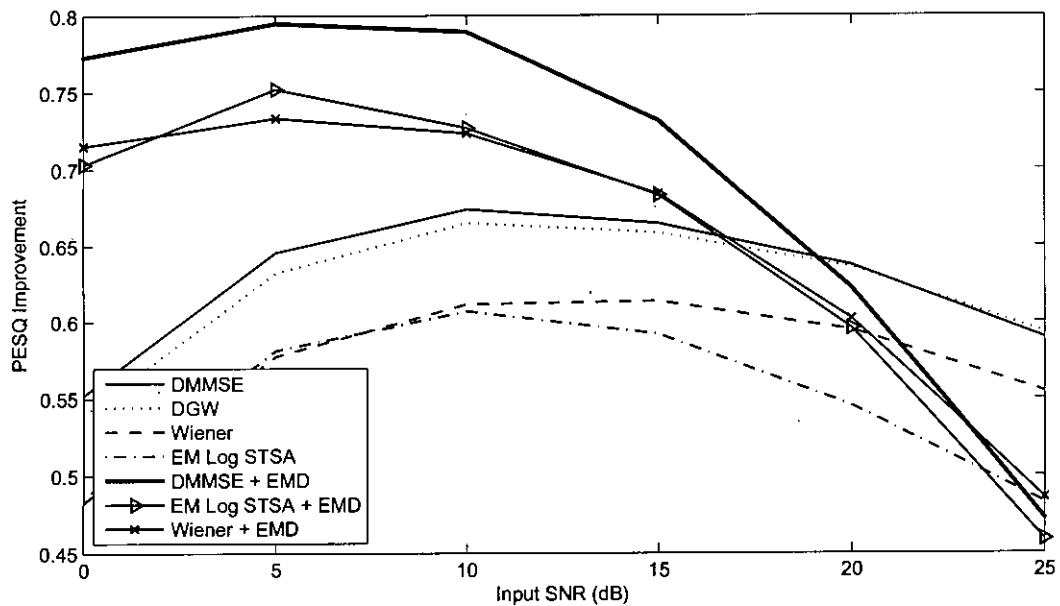


(b) Results using a polarity estimator of 80% accuracy.

Fig. 5.2: Performance comparison of the DMMSE, DGW, Wiener and EM Log STSA with respect to improvement in Overall SNR for an input SNR range of 0 dB to 25 dB. The proposed post filtering method is applied to the DMMSE, Wiener and EM Log STSA methods indicated by the +EMD notation. DMMSE+EMD indicates the proposed hybrid method. A polarity estimator of accuracy 100% and 80% was used in (a) and (b), respectively.



(a) Results using a polarity estimator of 100% accuracy (The ideal case).



(b) Results using a polarity estimator of 80% accuracy.

Fig. 5.3: Performance comparison of the DMMSE, DGW, Wiener and EM Log STSA with respect to improvement in PESQ scores for an input SNR range of 0 dB to 25 dB. The proposed post filtering method is applied to the DMMSE, Wiener and EM Log STSA methods indicated by the +EMD notation. DMMSE+EMD indicates the proposed hybrid method. A polarity estimator of accuracy 100% and 80% was used in (a) and (b), respectively.

Chapter 6

Conclusion

6.1 Summary

In this thesis, we have presented and evaluated a two stage hybrid speech enhancement algorithm aiming at an improved noise reduction performance in the first stage followed by a suppression of the musical noise in the second stage. The first stage of this method considers both the constructive and destructive interference of noise DCT coefficient, obtaining a set of optimum estimators in these conditional events assuming the exact joint probability distribution in the Gaussian speech and Gaussian noise statistical model. The proposed estimator, termed as the dual MMSE estimator, is shown to demonstrate superior performance with respect to MSE and SNR value improvement compared to the previously reported dual gain Wiener estimator, where a linear MMSE estimator was assumed. The main attribute of the dual MMSE estimator is that it handles three distinct cases of speech and noise DCT coefficient, when (1) the noise decreases the signal coefficient in magnitude, (2) the noise increases the signal coefficient in magnitude and (3) the noise reverses the polarity of the clean signal coefficient. The gain tends to be (1) attenuating, (2) amplifying and (3) negative, , respectively, in these three cases. This concept of negative value of a suppression rule, which concerns with the polarity correction of the noisy DCT coefficient, is quite novel and in contrast with the conventional definition of such rule. Even though this property was also present in the dual gain Wiener filter [3], the implications of this property was not properly addressed by the authors.

Towards accomplishing the second objective of the work, the newly developed empirical mode decomposition is utilized. A new post filtering technique is pro-

posed for suppression of residual noise remaining in the enhanced speech. The Chi-squared probability density function is assumed for a short duration of an intrinsic mode function energy and an optimum gain function is derived in the assumed model for residual noise suppression. The method is applied on the noisy IMF frames, each containing an integer number of cycles, in the time-domain. It is also observed that incorporating the speech presence uncertainty in the suppression rule provides more effective suppression of the unwanted noise. A novel method of noise variance estimation in the EMD domain is also presented using the energy-period relation of an IMF [31]. The proposed second stage algorithm is applied on speech files enhanced using traditional Wiener filter incorporating the variable averaging parameter [27] and the MMSE log spectral amplitude estimator [8]. The quality of the further enhanced speech has been evaluated using both subjective and objective quality measures. The improvement obtained in the average segmental SNR and PESQ values is remarkable. Even though these indices are known to be highly correlated to human listening, an actual listening test is also performed. The improvement of the mean opinion scores (MOS) and positive values in the comparative category rating (CCR) scores also indicate superior performance of the proposed second stage algorithm compared to well establish speech enhancement methods. Though the post filtering technique can perform on any enhanced speech file independent of the method, we have combined it with our proposed DMMSE estimator to form the hybrid method. The performance of the hybrid method is also tested and compared to traditional speech enhancement techniques and significant improvement in different quality measures is observed. The improvement was most prominent in the average segmental SNR and PESQ values whereas the overall SNR values were practically unchanged.

6.2 Future works

An efficient polarity estimator algorithm is very important to successfully implementing the dual MMSE estimator. Since the estimators are derived assuming that the constructive and destructive events are known *a priori*, proper identification of the events is necessary for the method to perform well. This identification must be done observing the noisy speech only, which is quite a challenging task.

Since we dealt with only the optimum estimator in the given events, this algorithm was beyond the scope of this thesis. Developing an efficient method for this purpose can be a prosperous future work.

Speech enhancement using the empirical mode decomposition is also a promising area of research. The new decomposition has not been very successfully exploited for speech enhancement in the first stage till date. This is due to the unavailability of a proper mathematical framework for analyzing signals in the EMD domain. The IMFs of the clean and noisy speech are assumed to be Gaussian in the derivation of the proposed suppression rule, which implies that the speech samples are normally distributed in time domain. This assumption, though greatly reduced the mathematical complexity of the problem (enabling the incorporation of the Chi-square distribution for short-time IMF energy), is not very accurate. Future works may thus include proper modeling of the IMFs generated from EMD of speech signals. Obviously, a mathematical model of the decomposition itself would greatly alleviate the problem, which is still unavailable in literature.

105821

Bibliography

- [1] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1218–1234, July 2006.
- [2] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and hilbert spectrum for non-linear and non-stationary time series analysis," *Proc. Roy. Soc. London A*, vol. 454, pp. 903–995, 1998.
- [3] I. Soon and S. Koh, "Low distortion speech enhancement," *Vision, Image and Signal Processing, IEE Proceedings* -, vol. 147, no. 3, pp. 247–253, Jun 2000.
- [4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, Apr 1979.
- [5] J. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 26, no. 5, pp. 471–472, Oct 1978.
- [6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504–512, Jul 2001.
- [7] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 28, no. 2, pp. 137–145, Apr 1980.

- [8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, Apr 1985.
- [9] —, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator."
- [10] P. Wolfe and S. Godsill, "Simple alternatives to the ephraim and malah suppression rule for speech enhancement," *Statistical Signal Processing, 2001. Proceedings of the 11th IEEE Signal Processing Workshop on*, pp. 496–499, 2001.
- [11] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2, pp. 629–632 vol. 2, 7-10 May 1996.
- [12] O. Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 345–349, Apr 1994.
- [13] Y. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *Signal Processing, IEEE Transactions on*, vol. 39, no. 9, pp. 1943–1954, Sep 1991.
- [14] J. Hansen and S. Nandkumar, "Robust estimation of speech in noisy backgrounds based on aspects of the auditory process," 1995. [Online]. Available: citeseer.ist.psu.edu/article/hansen95robust.html
- [15] D. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 6, pp. 497–514, Nov 1997.
- [16] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 2, pp. 126–137, Mar 1999.
- [17] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct 1992.

- [18] J. Deller, J. H. L. Hansen, and J. Proakis, *Discrete Time Processing of Speech Signals*. Vol. 1, IEEE Press, 2nd Edition,, 2000.
- [19] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.
- [20] —, "All-pole modeling of degraded speech," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 26, no. 3, pp. 197–210, Jun 1978.
- [21] J. Hansen and M. Clements, "Constrained iterative speech enhancement with application to speech recognition," *Signal Processing, IEEE Transactions on*, vol. 39, no. 4, pp. 795–805, Apr 1991.
- [22] H. Buchholz, *The Confluent Hypergeometric Function*. New York:Springer-Verlag, 1969.
- [23] R. M. Gray and L. D. Davisson, *An Introduction to Statistical Signal Processing*. Cambridge University Press, 2004.
- [24] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech Commun.*, vol. 24, no. 3, pp. 249–257, 1998.
- [25] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. John Wiley & Sons Ltd, 2006.
- [26] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. New York: McGraw-Hill, Inc., 1991.
- [27] M. K. Hasan, S. Salahuddin, and M. R. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," *Signal Processing Letters, IEEE*, vol. 11, no. 4, pp. 450–453, April 2004.
- [28] P. Flandrin, P. Goncalves, and G. Rilling, "Detrending and denoising with empirical mode decompositions," in *EUSIPCO-04*, 2004, pp. 1581–1584.
- [29] Z. F. Liu, Z. P. Liao, and E. F. Sang, "Speech enhancement based on hilbert-huang transform," *Machine Learning and Cybernetics, 2005. Proceedings of*

- 2005 *International Conference on*, vol. 8, pp. 4908–4912 Vol. 8, 18-21 Aug. 2005.
- [30] X. Zou, X. Li, and R. Zhang, “Speech enhancement based on hilbert-huang transform theory,” *Computer and Computational Sciences, 2006. IMSCCS '06. First International Multi-Symposiums on*, vol. 1, pp. 208–213, 20-24 June 2006.
- [31] Z. Wu and N. E. Huang *et. al.*, “A study of the characteristics of white noise using the empirical mode decomposition method,” *Proc. Roy. Soc. London A*, vol. 460, pp. 1597–1611, 2004.
- [32] J. Jensen, I. Batina, R. C. Hendriks, and R. Heusdens, “A study of the distribution of time-domain speech samples and discrete fourier coefficients,” *Proc. SPS-DARTS.*, vol. 1, pp. 155–158, 2005.
- [33] M. R. Islam, H. Haque, M. Q. Apu, and M. K. Hasan, “On the estimation of noise from pause regions for speech enhancement using spectral subtraction,” *International Conference on Electrical and Computer Engineering, 2004. ICECE'04.*, pp. 402–405, 28-30 December 2004.
- [34] Y. Hu and P. C. Loizou, “Evaluation of objective measures for speech enhancement,” *Proc. INTERSPEECH*, September 2006.
- [35] *Methods for subjective determination of transmission quality Annex E*, ITU-T Rec. P.800, 1996.
- [36] 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,” in *ICASSP '01: Proceedings of the Acoustics, Speech, and Signal Processing, 2001. on IEEE International Conference*. Washington, DC, USA: IEEE Computer Society, 2001, pp. 749–752.
- [37] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series and products*. New York: Academic Press, 1980, 5th corr. and enl. ed., 1980.

Appendix A

Important derivations

A.1 Derivation of the dual gain Wiener

Let $\mathbf{x}[n]$, $\mathbf{d}[n]$ and $\mathbf{y}[n]$ denote vectors containing the N most recent samples of the clean signal, noise and noisy signal, respectively, where N is the analysis frame size. If it is assumed that the noise is additive, $\mathbf{y}[n]$ can be expressed as

$$\mathbf{y}[n] = \mathbf{x}[n] + \mathbf{d}[n]. \quad (\text{A.1})$$

The DCT domain representation of (A.31) in the k -th frequency bin is

$$Y_k = X_k + D_k. \quad (\text{A.2})$$

If \hat{X}_k is an estimate of X_k , the MSE is given by,

$$J = E\{(\hat{X}_k - X_k)^2\}.$$

Assuming a Linear MMSE estimator such that $\hat{X}_k = WY_k$, we obtain the well known Wiener MSE given by,

$$J_W = E\{(WY_k - X_k)^2\} \quad (\text{A.3})$$

Substituting (A.32) into (A.3)

$$\begin{aligned} J_W &= E\{[W(X_k + D_k) - X_k]^2\} \\ &= (W^2 - 2W + 1)E\{X_k^2\} + 2W(W - 1)E\{X_k D_k\} \\ &\quad + W^2 E\{D_k^2\} \end{aligned} \quad (\text{A.4})$$

Minimizing (A.3) with respect to W leads to the conventional Wiener estimator, as shown in Section 2.3.1. In that derivation, $E\{X_k D_k\} = 0$ was assumed. To

consider the constructive and destructive events separately, let us define two mutually exclusive events as,

$$H_+: \quad \text{signal and noise are constructive: } X_k D_k \geq 0,$$

$$H_-: \quad \text{signal and noise are destructive: } X_k D_k < 0.$$

Now, if we want to derive a set of Wiener filters given these events have occurred, we shall obtain a pair of conditional MSEs, given by,

$$J_{W_+} = E\{(W_+ Y_k - X_k)^2 | H_+\} \quad (\text{A.5})$$

$$J_{W_-} = E\{(W_- Y_k - X_k)^2 | H_-\} \quad (\text{A.6})$$

Now, the terms J_{W_+} and J_{W_-} can be expanded as in (A.4) which will include the terms $E\{X_k^2 | H_+\}$, $E\{X_k^2 | H_-\}$, $E\{D_k^2 | H_+\}$, $E\{D_k^2 | H_-\}$, $E\{X_k D_k | H_+\}$ and $E\{X_k D_k | H_-\}$. Clearly, $E\{X_k^2 | H_+\} = E\{X_k^2 | H_-\} = E\{X_k^2\}$ and $E\{D_k^2 | H_+\} = E\{D_k^2 | H_-\} = E\{D_k^2\}$, since X_k and D_k are uncorrelated. But the cross terms need to be determined. Now, for the event H_+ , i.e., $X_k D_k > 0$

$$\begin{aligned} E\{X_k D_k | H_+\} &= E\{|X_k| |D_k|\} \\ &= E\{|X_k|\} E\{|D_k|\} \end{aligned}$$

and, for the event H_- , i.e., $X_k D_k < 0$

$$\begin{aligned} E\{X_k D_k | H_-\} &= -E\{|X_k| |D_k|\} \\ &= -E\{|X_k|\} E\{|D_k|\} \end{aligned}$$

Thus the cross terms $E\{X_k D_k | H_+\}$ and $E\{X_k D_k | H_-\}$ cannot be assumed to be zero in any of these events. To determine their values, first let us calculate the values of $E\{|X_k|\}$ and $E\{|D_k|\}$ assuming the Gaussian statistical model.

The probability density function of a random variable x which follows Gaussian distribution is

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp \left[-\frac{(x - \mu)^2}{2\sigma_x^2} \right], \quad -\infty < x < \infty \quad (\text{A.7})$$

where σ and μ are the mean and standard deviation of x , respectively. The expected value of x with the distribution given above is defined as

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \quad (\text{A.8})$$

As X_k is also a random variable and zero-mean ($\mu = 0$), the probability density function of X_k

$$f(X) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp(-X^2/2\sigma_x^2), -\infty < X < \infty \quad (\text{A.9})$$

But the probability density function of $|X_k|$ is required. A fundamental theorem on probability density function of a random variable y , when $y = g(x)$, is

$$f_y(y) = \frac{f_x(x_1)}{|g'(x_1)|} + \dots + \frac{f_x(x_n)}{|g'(x_n)|} + \dots \quad (\text{A.10})$$

where $g'(x)$ is the derivative of $g(x)$, $f_x(x)$ is the distribution of x and x_1, x_2, \dots, x_n are the real roots of $y = g(x)$.

In this case, $g(X) = |X|$, $y = g(x)$, i.e., $|X| = g(X)$ has two roots $+X$ and $-X$, $X_1 = +X$, $X_2 = -X$. Using (A.10)

$$f(|X|) = \frac{f_X(X_1)}{|g'(X_1)|} + \frac{f_X(X_2)}{|g'(X_2)|} \quad (\text{A.11})$$

As $g(X_1) = +X$ therefore $g'(X_1) = +1$ and $|g'(X_1)| = 1$. Similarly $g(X_2) = -X$ therefore $g'(X_2) = -1$ and $|g'(X_2)| = 1$. Also

$$\begin{aligned} f_X(X_1) &= \frac{1}{\sigma_x \sqrt{2\pi}} \exp(-(+X)^2/2\sigma_x^2) \\ &= \frac{1}{\sigma_x \sqrt{2\pi}} \exp(-X^2/2\sigma_x^2) \end{aligned} \quad (\text{A.12})$$

$$\begin{aligned} f_X(X_2) &= \frac{1}{\sigma_x \sqrt{2\pi}} \exp(-(-X)^2/2\sigma_x^2) \\ &= \frac{1}{\sigma_x \sqrt{2\pi}} \exp(-X^2/2\sigma_x^2) \end{aligned} \quad (\text{A.13})$$

Substituting Eqs. (A.12) and (A.13) into (A.11)

$$\begin{aligned} f(|X|) &= \frac{\frac{1}{\sigma_x \sqrt{2\pi}} \exp(-X^2/2\sigma_x^2)}{1} + \frac{\frac{1}{\sigma_x \sqrt{2\pi}} \exp(-X^2/2\sigma_x^2)}{1} \\ &= \frac{1}{\sigma_x \sqrt{2\pi}} \exp(-X^2/2\sigma_x^2) + \frac{1}{\sigma_x \sqrt{2\pi}} \exp(-X^2/2\sigma_x^2) \\ &= \frac{2}{\sigma_x \sqrt{2\pi}} \exp(-X^2/2\sigma_x^2) \end{aligned} \quad (\text{A.14})$$

The probability density function $f(|X|)$ is obtained as

$$f(|X|) = \frac{2}{\sigma_x \sqrt{2\pi}} \exp(-X^2/2\sigma_x^2) \quad (\text{A.15})$$

With the distribution function given in (A.15) and the definition of expected value given in (A.8), we get

$$\begin{aligned}
 E(|X|) &= \int_0^{\infty} X f(|X|) dX \\
 &= \int_0^{\infty} X \frac{2}{\sigma_x \sqrt{2\pi}} \exp(-X^2/2\sigma_x^2) dX \\
 &= \frac{2}{\sigma_x \sqrt{2\pi}} \int_0^{\infty} X \exp(-X^2/2\sigma_x^2) dX \quad (A.16)
 \end{aligned}$$

Now, using the formula

$$\int_0^{\infty} x^m \exp(-ax^2) dx = \frac{\Gamma(\frac{m+1}{2})}{2a^{\frac{m+1}{2}}}, \quad a > 0, \quad m > -1 \quad (A.17)$$

we obtain,

$$\int_0^{\infty} X \exp(-X^2/2\sigma_x^2) dX = \frac{\Gamma(\frac{1+1}{2})}{2(1/2\sigma_x^2)^{\frac{1+1}{2}}} \quad (A.18)$$

Thus $E\{|X|\}$ is obtained as

$$\begin{aligned}
 E(|X|) &= \frac{2}{\sigma_x \sqrt{2\pi}} \frac{\Gamma(\frac{1+1}{2})}{2(1/2\sigma_x^2)^{\frac{1+1}{2}}} \\
 &= \frac{2}{\sigma_x \sqrt{2\pi}} \frac{\Gamma(1)}{2(1/2\sigma_x^2)^1} \\
 &= \frac{2}{\sigma_x \sqrt{2\pi}} \frac{1}{2(1/2\sigma_x^2)} \\
 &= \sqrt{\frac{2}{\pi}} \sigma_x \quad (A.19)
 \end{aligned}$$

i.e.,

$$E\{|X_k|\} = \sqrt{\frac{2}{\pi}} \sigma_x \quad (A.20)$$

Similarly,

$$E(|D_k|) = \sqrt{\frac{2}{\pi}} \sigma_d \quad (A.21)$$

Finally, substituting Eqs. (A.20) and (A.21) into Eqs. (A.7) and (A.7), respectively,

$$E\{X_k D_k | H_+\} = \frac{2}{\pi} \sigma_x \sigma_d, \quad (A.22)$$

and

$$E\{X_k D_k | H_-\} = -\frac{2}{\pi} \sigma_x \sigma_d, \quad (A.23)$$

where σ_x and σ_d are the standard deviations of the clean speech spectral X_k and the noise spectral component D_k , respectively (i.e., $\sigma_x/\sigma_d = \sqrt{\xi_k}$). Expanding (A.5) as (A.4) and substituting (A.22) into the equation, we obtain

$$\begin{aligned} J_{W_+} &= (G_{W_+}^2 - 2G_{W_+} + 1)E\{X_k^2|H_+\} + 2G_{W_+}(G_{W_+} - 1)E\{X_k D_k|H_+\} \\ &\quad + G_{W_+}^2 E\{D_k^2|H_+\} \\ &= (G_{W_+}^2 - 2G_{W_+} + 1)E\{X_k^2\} + 2G_{W_+}(G_{W_+} - 1)\frac{2}{\pi}\sigma_x\sigma_d \\ &\quad + G_{W_+}^2 E\{D_k^2\} \end{aligned} \quad (\text{A.24})$$

Note that the gain is now defined as G_{W_+} instead of W_+ . Differentiating J_{W_+} with respect to G_{W_+} gives

$$\begin{aligned} \frac{\partial J_{W_+}}{\partial G_{W_+}} &= (2G_{W_+} - 2)E\{X_k^2\} + 2(2G_{W_+} - 1)\frac{2}{\pi}\sigma_x\sigma_d \\ &\quad + 2G_{W_+}E\{D_k^2\} \end{aligned} \quad (\text{A.25})$$

Equating $\partial J_{W_+}/\partial G_{W_+}$ to zero yields

$$(2G_{W_+} - 2)E\{X_k^2\} + 2(2G_{W_+} - 1)\frac{2}{\pi}\sigma_x\sigma_d + 2G_{W_+}E\{D_k^2\} = 0 \quad (\text{A.26})$$

Dividing (A.26) by $2E\{D_k^2\}$ and substituting $E\{X_k^2\}/E\{D_k^2\} = \xi_k$ (defined in (2.17))

$$(G_{W_+} - 1)\xi_k + (2G_{W_+} - 1)\frac{\frac{2}{\pi}\sigma_x\sigma_d}{E\{D_k^2\}} + G_{W_+} = 0 \quad (\text{A.27})$$

Substituting $E\{D_k^2\} = \sigma_d^2$ in (A.27), we obtain

$$\begin{aligned} (G_{W_+} - 1)\xi_k + (2G_{W_+} - 1)\frac{2}{\pi}\frac{\sigma_x}{\sigma_d} + G_{W_+} &= (G_{W_+} - 1)\xi_k \\ &\quad + (2G_{W_+} - 1)\frac{2}{\pi}\sqrt{\xi_k} + G_{W_+} \\ &= G_{W_+}(\xi_k + 1 + \frac{4}{\pi}\sqrt{\xi_k}) \\ &\quad - \xi_k - \frac{2}{\pi}\sqrt{\xi_k} \end{aligned} \quad (\text{A.28})$$

Rearranging (A.28), the optimum filter gain in the event H_+ ,

$$G_{W_+} = \frac{\xi_k + \frac{2}{\pi}\sqrt{\xi_k}}{\xi_k + 1 + \frac{4}{\pi}\sqrt{\xi_k}} \quad (\text{A.29})$$

G_{W_+} is always less than 1 and, the authors [3] have proposed to use this gain for the spectral component whose magnitude has been increased by noise, i.e., for the condition $X_k D_k > 0$.

Similarly, substituting (A.23) into (A.6) and equating $\partial J_{W-}/\partial G_{W-}$ to zero gives the optimum filter gain in the event H_- as,

$$G_{W-} = \frac{\xi_k - \frac{2}{\pi}\sqrt{\xi_k}}{\xi_k + 1 - \frac{4}{\pi}\sqrt{\xi_k}} \quad (\text{A.30})$$

The authors [3] have proposed to use G_{W+} for the spectral component whose magnitude has been reduced by noise, i.e., for the condition $X_k D_k < 0$.

A.2 MMSE estimator in the DCT domain

Let $\mathbf{x}[n]$, $\mathbf{d}[n]$ and $\mathbf{y}[n]$ denote vectors containing the N most recent samples of the clean signal, noise and noisy signal, respectively, where N is the analysis frame size. If it is assumed that the noise is additive, $\mathbf{y}[n]$ can be expressed as

$$\mathbf{y}[n] = \mathbf{x}[n] + \mathbf{d}[n]. \quad (\text{A.31})$$

The DCT domain representation of (A.31) in the k -th frequency bin is

$$Y_k = X_k + D_k. \quad (\text{A.32})$$

With the assumption that the DCT transform coefficients are statistically independent, the Minimum Mean Square Error (MMSE) estimated amplitude \hat{X}_k can be obtained from Y_k as follows:

$$\hat{X}_k = E\{X_k|Y_k\} \quad (\text{A.33})$$

Where $E\{\cdot\}$ denotes the expectation operator. (A.33) may be expressed as,

$$\begin{aligned} \hat{X}_k &= \int_{-\infty}^{\infty} x_k p(x_k|Y_k) dx_k \\ &= \frac{\int_{-\infty}^{\infty} x_k p(x_k, Y_k) dx_k}{p(Y_k)} \end{aligned} \quad (\text{A.34})$$

Now the Gaussian statistical model assumption leads to the following marginal and joint distributions:

$$p(x_k) = \frac{1}{2\pi\sigma_x^2} \exp\left[-\frac{x_k^2}{2\sigma_x^2}\right] \quad (\text{A.35})$$

$$p_{XY}(x_k, y_k) = \frac{1}{2\pi\sigma_d\sigma_x} \exp\left[-\frac{x_k^2}{2\sigma_x^2} - \frac{(y_k - x_k)^2}{2\sigma_d^2}\right] \quad (\text{A.36})$$

Where σ_x^2 and σ_d^2 denote the variances of the k th spectral components of the clean signal and noise processes. To evaluate (A.34), let,

$$\begin{aligned} I_1(a, b) &= \int_a^b x_k p(x_k, Y_k) dx_k \quad (\text{A.37}) \\ &= \int_a^b \frac{x_k}{2\pi\sigma_d\sigma_x} \exp\left[-\frac{x_k^2}{2\sigma_x^2} - \frac{(x_k - Y_k)^2}{2\sigma_d^2}\right] dx_k \\ &= \frac{\exp\left(-\frac{Y_k^2}{2\sigma_y^2}\right)}{2\pi\sigma_d\sigma_x} \int_a^b x_k \exp\left[-\frac{1}{2}\left(\frac{x_k\sigma_y}{\sigma_x\sigma_d} - \frac{\sigma_x Y_k}{\sigma_y\sigma_d}\right)^2\right] dx_k \end{aligned}$$

Where, $\sigma_y = \sqrt{\sigma_x^2 + \sigma_d^2}$. Letting $z = \frac{1}{\sqrt{2}}\left(\frac{x_k\sigma_y}{\sigma_x\sigma_d} - \frac{\sigma_x Y_k}{\sigma_y\sigma_d}\right)$

$$I_1(a, b) = \frac{\exp\left(-\frac{Y_k^2}{2\sigma_y^2}\right)}{2\pi\sigma_d\sigma_x} \int_{a_1}^{b_1} \left[z\left(\frac{\sqrt{2}\sigma_x\sigma_d}{\sigma_y}\right) + \left(\frac{\sigma_x^2 Y_k}{\sigma_y^2}\right)\right] e^{-z^2} \left(\frac{\sqrt{2}\sigma_x\sigma_d}{\sigma_y}\right) dz$$

Where,

$$a_1 = \frac{1}{\sqrt{2}}\left(a\frac{\sigma_y}{\sigma_x\sigma_d} - \frac{\sigma_x Y_k}{\sigma_y\sigma_d}\right) \quad (\text{A.38})$$

$$b_1 = \frac{1}{\sqrt{2}}\left(b\frac{\sigma_y}{\sigma_x\sigma_d} - \frac{\sigma_x Y_k}{\sigma_y\sigma_d}\right) \quad (\text{A.39})$$

Thus,

$$\begin{aligned} I_1(a, b) &= \frac{e^{-\frac{Y_k^2}{2\sigma_y^2}}}{\sqrt{2}\pi\sigma_y} \left[\frac{\sqrt{2}\sigma_x\sigma_d}{\sigma_y} \int_{a_1}^{b_1} z e^{-z^2} dz + \frac{\sigma_x^2 Y_k}{\sigma_y^2} \int_{a_1}^{b_1} e^{-z^2} dz \right] \\ &= \frac{e^{-\frac{Y_k^2}{2\sigma_y^2}}}{\sqrt{2}\pi\sigma_y} \left[\frac{\sqrt{2}\sigma_x\sigma_d}{\sigma_y} \frac{1}{2} (e^{-a_1^2} - e^{-b_1^2}) + \frac{\sigma_x^2 Y_k}{\sigma_y^2} \frac{\sqrt{\pi}}{2} (\text{erf}(b_1) - \text{erf}(a_1)) \right] \\ &= \frac{e^{-\frac{Y_k^2}{2\sigma_y^2}}}{\sqrt{2}\pi\sigma_y} \left[\frac{\sigma_x\sigma_d}{\sqrt{2}\pi\sigma_y} (e^{-a_1^2} - e^{-b_1^2}) + \frac{\sigma_x^2 Y_k}{2\sigma_y^2} (\text{erf}(b_1) - \text{erf}(a_1)) \right] \\ &= p(Y_k) \left[\frac{\sigma_x\sigma_d}{\sqrt{2}\pi\sigma_y} (e^{-a_1^2} - e^{-b_1^2}) + \frac{\sigma_x^2 Y_k}{2\sigma_y^2} (\text{erf}(b_1) - \text{erf}(a_1)) \right] \quad (\text{A.40}) \end{aligned}$$

Where,

$$p(Y_k) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{Y_k^2}{2\sigma_y^2}\right) \quad (\text{A.41})$$

and,

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \quad (\text{A.42})$$

Now, in (A.34), we have the limits $a = -\infty$ and $b = \infty$ for $I_1(a, b)$ as defined in (A.49), which also gives $a_1 = -\infty$ and $b_1 = \infty$. Thus, using (A.40) we have from (A.34),

$$\begin{aligned}
\hat{X}_k &= \frac{p(Y_k) \frac{\sigma_x^2 Y_k}{2\sigma_y^2} (\text{erf}(\infty) - \text{erf}(-\infty))}{p(Y_k)} \\
&= \frac{\sigma_x^2 Y_k}{\sigma_y^2} \\
&= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_d^2} Y_k
\end{aligned} \tag{A.43}$$

This is the conventional Wiener estimator which can also be derived assuming a linear MMSE estimator¹. Since X_k and Y_k are jointly Gaussian, the optimal MMSE in this case linear. The Wiener estimator can be viewed as a gain multiplied by the noisy coefficient as,

$$\hat{X}_k^W = G_W \times Y_k \tag{A.44}$$

where, the Wiener gain function,

$$\begin{aligned}
G_W &= \frac{\xi_k}{1 + \xi_k}, \\
\text{and } \xi_k &= \frac{E\{X_k^2\}}{E\{D_k^2\}}
\end{aligned} \tag{A.45}$$

is interpreted as the *a priori* SNR [7].

A.3 Derivation of the dual MMSE estimators

A.3.1 The marginal densities $p(Y_k, H_+)$ and $p(Y_k, H_-)$

We define two mutually exclusive events as:

$$\begin{aligned}
H_+ &: \quad \text{signal and noise are constructive: } X_k D_k \geq 0 \\
H_- &: \quad \text{signal and noise are destructive: } X_k D_k < 0
\end{aligned}$$

¹That is, assuming $\hat{X}_k = WY_k$ and minimizing $E\{(X_k - \hat{X}_k)^2\}$ for W .

Since Y_k is constructed from the two mutually exclusive events H_+ and H_- , we may write

$$\begin{aligned} E\{X_k|Y_k, H_+\} &= \int_{-\infty}^{\infty} x_k p(x_k|Y_k, H_+) dx_k \\ &= \int_{-\infty}^{\infty} x_k \frac{p(x_k, Y_k, H_+)}{p(Y_k, H_+)} dx_k \end{aligned} \quad (\text{A.46})$$

$$\begin{aligned} E\{X_k|Y_k, H_-\} &= \int_{-\infty}^{\infty} x_k p(x_k|Y_k, H_-) dx_k \\ &= \int_{-\infty}^{\infty} x_k \frac{p(x_k, Y_k, H_-)}{p(Y_k, H_-)} dx_k \end{aligned} \quad (\text{A.47})$$

Thus the quantities in (A.46) and (A.47) can be determined if we know the joint density functions $p(x_k, y_k, H_+)$, $p(x_k, y_k, H_-)$, $p(Y_k, H_+)$ and $p(Y_k, H_-)$.

We note that, for the event H_+ , $X_k D_k > 0$, which results in $X_k Y_k > X_k^2$ using (A.32). This condition simplifies to $m_k Y_k > |X_k|$ where $m_k = \text{sgn}(X_k)$. Similarly, the event H_- results in the condition $m_k Y_k < |X_k|$. These constraints of X_k and Y_k are shown graphically in Fig. 3.1 (a).

Thus the joint density functions $p(x_k, y_k, H_+)$ and $p(x_k, y_k, H_-)$ will be defined as,

$$\begin{aligned} p(x_k, y_k, H_+) &= \begin{cases} p_{XY}(x_k, y_k) & m_k Y_k > |X_k| \\ 0 & \text{otherwise} \end{cases} \\ p(x_k, y_k, H_-) &= \begin{cases} p_{XY}(x_k, y_k) & m_k Y_k < |X_k| \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

To evaluate (A.46) and (A.47), we now need to find the probability densities $p(Y_k, H_+)$ and $p(Y_k, H_-)$. We have,

$$p(Y_k, H_+) = \int_{-\infty}^{\infty} p_{XY}(x_k, Y_k, H_+) dx_k$$

We note that this integration must be solved separately for positive and negative values of Y_k .

$$\begin{aligned} p(Y_k, H_+) &= \begin{cases} \int_0^{Y_k} p_{XY}(x_k, Y_k) dx_k & \text{if } Y_k \geq 0 \\ \int_{Y_k}^0 p_{XY}(x_k, Y_k) dx_k & \text{if } Y_k < 0 \end{cases} \\ &= \text{sgn}(Y_k) \int_0^{Y_k} p_{XY}(x_k, Y_k) dx_k \end{aligned} \quad (\text{A.48})$$

To evaluate this integral we let,

$$\begin{aligned}
I_2(a, b) &= \int_a^b p(x_k, Y_k) dx_k & (A.49) \\
&= \int_a^b \frac{1}{2\pi\sigma_d\sigma_x} \exp\left[-\frac{x_k^2}{2\sigma_x^2} - \frac{(x_k - Y_k)^2}{2\sigma_d^2}\right] dx_k \\
&= \frac{\exp\left(-\frac{Y_k^2}{2\sigma_y^2}\right)}{2\pi\sigma_d\sigma_x} \int_a^b \exp\left[-\frac{1}{2}\left(\frac{x_k\sigma_y}{\sigma_x\sigma_d} - \frac{\sigma_x Y_k}{\sigma_y\sigma_d}\right)^2\right] dx_k
\end{aligned}$$

Again letting $z = \frac{1}{\sqrt{2}}\left(\frac{x_k\sigma_y}{\sigma_x\sigma_d} - \frac{\sigma_x Y_k}{\sigma_y\sigma_d}\right)$

$$I_2(a, b) = \frac{\exp\left(-\frac{Y_k^2}{2\sigma_y^2}\right)}{2\pi\sigma_d\sigma_x} \int_{a_1}^{b_1} e^{-z^2} \left(\frac{\sqrt{2}\sigma_x\sigma_d}{\sigma_y}\right) dz$$

Where,

$$a_1 = \frac{1}{\sqrt{2}}\left(a\frac{\sigma_y}{\sigma_x\sigma_d} - \frac{\sigma_x Y_k}{\sigma_y\sigma_d}\right) \quad (A.50)$$

$$b_1 = \frac{1}{\sqrt{2}}\left(b\frac{\sigma_y}{\sigma_x\sigma_d} - \frac{\sigma_x Y_k}{\sigma_y\sigma_d}\right) \quad (A.51)$$

Thus,

$$\begin{aligned}
I_2(a, b) &= \frac{e^{-\frac{Y_k^2}{2\sigma_y^2}}}{\sqrt{2}\pi\sigma_y} \int_{a_1}^{b_1} e^{-z^2} dz \\
&= \frac{e^{-\frac{Y_k^2}{2\sigma_y^2}}}{\sqrt{2}\pi\sigma_y} \frac{\sqrt{\pi}}{2} [\text{erf}(b_1) - \text{erf}(a_1)] \\
&= \frac{p(Y_k)}{2} [\text{erf}(b_1) - \text{erf}(a_1)]
\end{aligned}$$

Now, to evaluate $p(Y_k, H_+)$ we simply need to put the limits $a = 0$ and $b = Y_k$ which leads to $a_1 = -\frac{Y_k\sigma_x}{\sqrt{2}\sigma_y\sigma_d}$ and $b_1 = \frac{Y_k\sigma_d}{\sqrt{2}\sigma_x\sigma_y}$. Putting these values in (A.52),

$$p(Y_k, H_+) = \frac{\text{sgn}(Y_k)p(Y_k)}{2} \left[\text{erf}\left(\frac{Y_k\sigma_d}{\sqrt{2}\sigma_x\sigma_y}\right) + \text{erf}\left(\frac{Y_k\sigma_x}{\sqrt{2}\sigma_y\sigma_d}\right) \right]$$

Similarly, it can be shown that,

$$p(Y_k, H_-) = \begin{cases} \int_{-\infty}^0 p_{XY}(x_k, Y_k) dx_k + \int_{Y_k}^{\infty} p_{XY}(x_k, Y_k) dx_k & \text{if } Y_k \geq 0 \\ \int_{Y_k}^{\infty} p_{XY}(x_k, Y_k) dx_k + \int_{-\infty}^0 p_{XY}(x_k, Y_k) dx_k & \text{if } Y_k < 0 \end{cases}$$

Since $p_{XY}(x_k, y_k)$ is a symmetric function with respect to x_k and y_k , we may write,

$$\begin{aligned}
p(Y_k, H_-) &= \text{sgn}(Y_k) \left[\int_{-\infty}^0 p_{XY}(x_k, Y_k) dx_k + \int_{Y_k}^{\infty} p_{XY}(x_k, Y_k) dx_k \right] \\
&= \text{sgn}(Y_k) [I_2(-\infty, 0) + I_2(Y_k, \infty)] \\
&= \frac{\text{sgn}(Y_k)p(Y_k)}{2} \left[-\text{erf}(-\infty) + \text{erf}\left(-\frac{Y_k\sigma_y}{\sqrt{2}\sigma_x\sigma_d}\right) \right] \\
&\quad + \frac{\text{sgn}(Y_k)p(Y_k)}{2} \left[-\text{erf}\left(\frac{Y_k\sigma_d}{\sqrt{2}\sigma_x\sigma_y}\right) + \text{erf}(\infty) \right] \\
&= \frac{\text{sgn}(Y_k)p(Y_k)}{2} \left[1 - \text{erf}\left(-\frac{Y_k\sigma_y}{\sqrt{2}\sigma_x\sigma_d}\right) - \text{erf}\left(-\frac{Y_k\sigma_d}{\sqrt{2}\sigma_x\sigma_y}\right) + 1 \right] \\
&= \frac{\text{sgn}(Y_k)p(Y_k)}{2} \left[\text{erfc}\left(\frac{Y_k\sigma_y}{\sqrt{2}\sigma_x\sigma_d}\right) + \text{erfc}\left(\frac{Y_k\sigma_d}{\sqrt{2}\sigma_x\sigma_y}\right) \right] \quad (\text{A.52})
\end{aligned}$$

Here, $\text{erfc}(\cdot)$ is the complementary error function defined as,

$$\text{erfc}(x) = 1 - \text{erf}(x) \quad (\text{A.53})$$

A.3.2 The dual MMSE estimators

From (A.46) we may write,

$$E\{X_k|Y_k, H_+\} = \frac{\int_{-\infty}^{\infty} x_k p(x_k, Y_k, H_+) dx_k}{p(Y_k, H_+)} \quad (\text{A.54})$$

We have,

$$\begin{aligned}
\int_{-\infty}^{\infty} x_k p(x_k, Y_k, H_+) dx_k &= \begin{cases} \int_0^{Y_k} x_k p(x_k, Y_k) dx_k & \text{if } Y_k \geq 0 \\ \int_{Y_k}^0 x_k p(x_k, Y_k) dx_k & \text{if } Y_k < 0 \end{cases} \\
&= \text{sgn}(Y_k) \int_0^{Y_k} p_{XY}(x_k, Y_k) dx_k \\
&= \text{sgn}(Y_k) I_1(0, Y_k) \quad (\text{A.55})
\end{aligned}$$

Putting $a = 0$ and $b = Y_k$ in (A.40),

$$I_1(0, Y_k) = p(Y_k) \left[\frac{\sigma_x\sigma_d}{\sqrt{2\pi}\sigma_y} (e^{-a_1^2} - e^{-b_1^2}) + \frac{\sigma_x^2 Y_k}{2\sigma_y^2} (\text{erf}(b_1) - \text{erf}(a_1)) \right] \quad (\text{A.56})$$

Here, $a_1 = -\frac{Y_k \sigma_x}{\sqrt{2} \sigma_y \sigma_d}$ and $b_1 = \frac{Y_k \sigma_d}{\sqrt{2} \sigma_x \sigma_y}$. Thus, from (A.54)

$$\begin{aligned} E\{X_k|Y_k, H_+\} &= \frac{\text{sgn}(Y_k)p(Y_k) \left[\frac{\sigma_x \sigma_d}{\sqrt{2\pi} \sigma_y} (e^{-a_1^2} - e^{-b_1^2}) + \frac{\sigma_x^2 Y_k}{2\sigma_y^2} \{\text{erf}(b_1) - \text{erf}(a_1)\} \right]}{\frac{\text{sgn}(Y_k)p(Y_k)}{2} [\text{erf}(b_1) - \text{erf}(a_1)]} \\ &= \frac{\sigma_x^2 Y_k}{\sigma_y^2} + \sqrt{\frac{2}{\pi}} \frac{\sigma_x \sigma_d}{\sigma_y} \left[\frac{e^{-a_1^2} - e^{-b_1^2}}{\text{erf}(b_1) - \text{erf}(a_1)} \right] \end{aligned} \quad (\text{A.57})$$

We know,

$$\sigma_y^2 = \sigma_x^2 + \sigma_d^2$$

and

$$G_W = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_d^2} = \frac{\sigma_x^2}{\sigma_y^2} \quad (\text{A.58})$$

Let,

$$\begin{aligned} f_1 &= b_1 = \frac{Y_k \sigma_d}{\sqrt{2} \sigma_x \sqrt{\sigma_d^2 + \sigma_x^2}} \\ f_2 &= -a_1 = \frac{Y_k \sigma_x}{\sqrt{2} \sigma_d \sqrt{\sigma_d^2 + \sigma_x^2}} \end{aligned}$$

Since $\text{erf}(\cdot)$ is an odd function, from (A.57),

$$\begin{aligned} E\{X_k|Y_k, H_+\} &= G_W Y_k + \sqrt{\frac{2}{\pi}} \left[\frac{\exp(-f_1^2) - \exp(-f_2^2)}{\text{erf}(f_1) + \text{erf}(f_2)} \right] \frac{\sigma_d \sigma_x}{\sqrt{\sigma_d^2 + \sigma_x^2}} \\ &= G_W Y_k + \Phi(Y_k) \end{aligned} \quad (\text{A.59})$$

where,

$$\Phi(Y_k) = \sqrt{\frac{2}{\pi}} \left[\frac{\exp(-f_1^2) - \exp(-f_2^2)}{\text{erf}(f_1) + \text{erf}(f_2)} \right] \frac{\sigma_d \sigma_x}{\sqrt{\sigma_d^2 + \sigma_x^2}}. \quad (\text{A.60})$$

Noting that $\Phi(Y_k)$, containing f_1 and f_2 , is an odd function, it can be expressed as

$$\Phi(Y_k) = \text{sgn}(Y_k) \Phi(|Y_k|),$$

enabling us to express (A.59) as a gain expression multiplied by the noisy DCT component, Y_k , i.e.,

$$\begin{aligned} E\{X_k|Y_k, H_+\} &= \left(G_W + \frac{\Phi(Y_k)}{Y_k} \right) Y_k \\ &= \left(G_W + \frac{\Phi(|Y_k|)}{|Y_k|} \right) Y_k \\ &= G_{\text{MMSE}^+} \times Y_k \end{aligned} \quad (\text{A.61})$$

where $G_{\text{MMSE}+}$ denotes the MMSE gain function for the event H_+ . Now, the *a priori* and *a posteriori* SNRs are defined as,

$$\xi_k = \frac{\sigma_x^2}{\sigma_d^2}, \quad \text{and}$$

$$\gamma_k = \frac{|Y_k|^2}{\sigma_d^2},$$

respectively. Expressing $\Phi(|Y_k|)$ using ξ_k and γ_k ,

$$\begin{aligned} \Phi(|Y_k|) &= \sqrt{\frac{2}{\pi}} \left[\frac{\exp\left(-\frac{|Y_k|^2 \sigma_d^2}{2\sigma_x^2(\sigma_d^2 + \sigma_x^2)}\right) - \exp\left(-\frac{|Y_k|^2 \sigma_x^2}{2\sigma_d^2(\sigma_d^2 + \sigma_x^2)}\right)}{\operatorname{erf}\left(\frac{|Y_k| \sigma_d}{\sqrt{2}\sigma_x \sqrt{\sigma_d^2 + \sigma_x^2}}\right) + \operatorname{erf}\left(\frac{|Y_k| \sigma_x}{\sqrt{2}\sigma_d \sqrt{\sigma_d^2 + \sigma_x^2}}\right)} \right] \frac{\sigma_d \sigma_x}{\sqrt{\sigma_d^2 + \sigma_x^2}} \\ &= \sigma_d \sqrt{\frac{2}{\pi}} \left[\frac{\exp\left(-\frac{\gamma_k}{2\xi_k(1+\xi_k)}\right) - \exp\left(-\frac{\gamma_k \xi_k}{2(1+\xi_k)}\right)}{\operatorname{erf}\left(\sqrt{\frac{\gamma_k}{2\xi_k(1+\xi_k)}}\right) + \operatorname{erf}\left(\sqrt{\frac{\gamma_k \xi_k}{2(1+\xi_k)}}\right)} \right] \sqrt{\frac{\xi_k}{1+\xi_k}} \quad (\text{A.62}) \end{aligned}$$

Thus,

$$\begin{aligned} G_{\text{MMSE}+} &= G_w + \frac{\Phi(|Y_k|)}{|Y_k|} \\ &= \frac{\xi_k}{\xi_k + 1} + \sqrt{\frac{\xi_k}{1+\xi_k}} \sqrt{\frac{2}{\pi \gamma_k}} \left[\frac{e^{-\frac{\gamma_k}{2\xi_k(1+\xi_k)}} - e^{-\frac{\gamma_k \xi_k}{2(1+\xi_k)}}}{\operatorname{erf}\left(\sqrt{\frac{\gamma_k}{2\xi_k(1+\xi_k)}}\right) + \operatorname{erf}\left(\sqrt{\frac{\gamma_k \xi_k}{2(1+\xi_k)}}\right)} \right]. \end{aligned}$$

Following a very similar method, the gain for the destructive case, $G_{\text{MMSE}-}$ can be found to be,

$$G_{\text{MMSE}-} = \frac{\xi_k}{\xi_k + 1} - \sqrt{\frac{\xi_k}{1+\xi_k}} \sqrt{\frac{2}{\pi \gamma_k}} \left[\frac{e^{-\frac{\gamma_k}{2\xi_k(1+\xi_k)}} - e^{-\frac{\gamma_k \xi_k}{2(1+\xi_k)}}}{\operatorname{erfc}\left(\sqrt{\frac{\gamma_k}{2\xi_k(1+\xi_k)}}\right) + \operatorname{erfc}\left(\sqrt{\frac{\gamma_k \xi_k}{2(1+\xi_k)}}\right)} \right].$$

A.4 The Ephraim and Malah suppression rule

Assuming the Gaussian Model, Ephraim and Malah [9] derived a minimum mean-square error (MMSE) short-time spectral amplitude estimator under the assumption that the Fourier expansion coefficients of the original signal and the noise may be modelled as statistically independent, zero-mean, Gaussian random variables. Thus the observed spectral component in the bin k , $\mathbf{Y}_k \triangleq R_k \exp(j\vartheta_k)$, is equal to the sum of the spectral components of the signal, $\mathbf{X}_k \triangleq A_k \exp(j\alpha_k)$, and the noise, \mathbf{D}_k . That is,

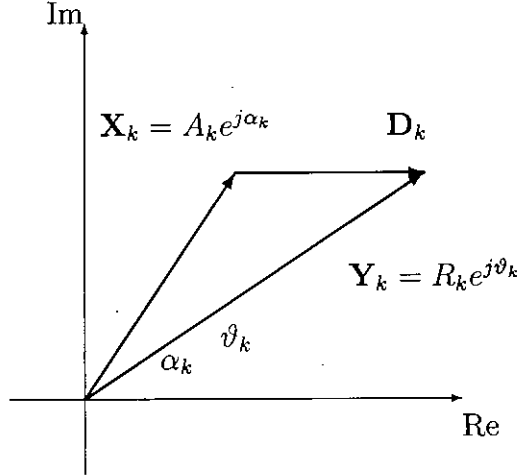


Fig. A.1: Vector Representation of $\mathbf{Y}_k = \mathbf{X}_k + \mathbf{D}_k$

$$\mathbf{Y}_k = \mathbf{X}_k + \mathbf{D}_k \quad (\text{A.63})$$

In this derivation, only the phase of \mathbf{Y}_k and \mathbf{X}_k are assumed, while the phase angle of \mathbf{D}_k is considered to be zero. This generalization simplifies the derivation without any loss of generality since the sum of all three phase angles must be equal to 2π .

It is obvious that the amplitude of a complex Gaussian random variable will follow a Rayleigh distribution [26]. Thus we may assume the following distributions for a_k and α_k :

$$p(a_k) = \begin{cases} \frac{2a_k}{\lambda_x(k)} \exp\left(-\frac{a_k^2}{\lambda_x(k)}\right) & \text{if } a_k \in [0, \infty), \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.64})$$

$$p(\alpha_k) = \begin{cases} \frac{1}{2\pi} & \text{if } \alpha_k \in [-\pi, \pi), \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.65})$$

$$(\text{A.66})$$

The joint PDF of a_k and α_k will be then given by:

$$p(a_k, \alpha_k) = \frac{a_k}{\pi \lambda_x(k)} \exp\left(-\frac{a_k^2}{\lambda_x(k)}\right) \quad (\text{A.67})$$

Now, the noise DFT coefficients is assumed to follow a complex Gaussian distribution. Thus:

$$p(\mathbf{D}_k) = \frac{1}{\pi \lambda_d(k)} \exp\left(-\frac{|\mathbf{D}_k|^2}{\lambda_d(k)}\right) \quad (\text{A.68})$$

From (A.63) it is obvious that $\mathbf{D}_k = \mathbf{Y}_k - a_k e^{j\alpha_k}$. Thus the conditional probability distribution of the noisy signal DFT coefficient will be given by:

$$p(\mathbf{Y}_k | a_k, \alpha_k) = \frac{1}{\pi \lambda_d(k)} \exp\left(-\frac{|\mathbf{Y}_k - a_k e^{j\alpha_k}|^2}{\lambda_d(k)}\right) \quad (\text{A.69})$$

Now the MMSE estimation problem is reduced to that of estimating A_k from the observations of \mathbf{Y}_k . This estimate \hat{A}_k is obtained as follows:

$$\begin{aligned} \hat{A}_k &= E\{A_k | \mathbf{Y}_k\} \\ &= \frac{\int_0^\infty \int_0^{2\pi} a_k p(\mathbf{Y}_k | a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k}{\int_0^\infty \int_0^{2\pi} p(\mathbf{Y}_k | a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k} \\ &= \frac{\int_0^\infty a_k^2 \exp\left(-\frac{a_k^2}{\lambda_x(k)}\right) \int_0^{2\pi} \exp\left(-\frac{|\mathbf{Y}_k - a_k e^{j\alpha_k}|^2}{\lambda_d(k)}\right) d\alpha_k da_k}{\int_0^\infty a_k \exp\left(-\frac{a_k^2}{\lambda_x(k)}\right) \int_0^{2\pi} \exp\left(-\frac{|\mathbf{Y}_k - a_k e^{j\alpha_k}|^2}{\lambda_d(k)}\right) d\alpha_k da_k} \end{aligned} \quad (\text{A.70})$$

Now, using the vector addition rule we may express the integral of α_k as,

$$\begin{aligned} \int_0^{2\pi} \exp\left(-\frac{|\mathbf{Y}_k - a_k e^{j\alpha_k}|^2}{\lambda_d(k)}\right) d\alpha_k &= \int_0^{2\pi} \exp\left(-\frac{R_k^2 + a_k^2 - 2a_k R_k \cos(\alpha_k - \vartheta_k)}{\lambda_d(k)}\right) d\alpha_k \\ &= \exp\left(-\frac{R_k^2 + a_k^2}{\lambda_d(k)}\right) \int_0^{2\pi} \exp\left(\frac{2a_k R_k \cos \beta}{\lambda_d(k)}\right) d\beta \end{aligned} \quad (\text{A.71})$$

Where $\beta = \alpha_k - \vartheta_k$ is assumed. From the Integral form of the Modified Bessel Function, we know,

$$I_n(z) = \int_0^{2\pi} \cos(\beta n) \exp(z \cos \beta) d\beta \quad (\text{A.72})$$

We note that the integral of (A.71) can be expressed as in (A.72) using $z = \frac{2a_k R_k}{\lambda_d}$ and $n = 0$. Thus we have from (A.71)

$$\int_0^{2\pi} \exp\left(-\frac{|\mathbf{Y}_k - a_k e^{j\alpha_k}|^2}{\lambda_d(k)}\right) d\alpha_k = \exp\left(-\frac{R_k^2 + a_k^2}{\lambda_d(k)}\right) I_0\left(\frac{2a_k R_k}{\lambda_d}\right) \quad (\text{A.73})$$

Substituting this expression in (A.70) we have,

$$\begin{aligned}
\hat{A}_k &= \frac{\int_0^\infty a_k^2 \exp\left(-\frac{a_k^2}{\lambda_x(k)}\right) \exp\left(-\frac{R_k^2 + a_k^2}{\lambda_d(k)}\right) I_0\left(\frac{2a_k R_k}{\lambda_d(k)}\right) da_k}{\int_0^\infty a_k \exp\left(-\frac{a_k^2}{\lambda_x(k)}\right) \exp\left(-\frac{R_k^2 + a_k^2}{\lambda_d(k)}\right) I_0\left(\frac{2a_k R_k}{\lambda_d(k)}\right) da_k} \\
&= \frac{\int_0^\infty a_k^2 \exp\left(-\frac{R_k^2}{\lambda_d(k)}\right) \exp\left(-\frac{a_k^2}{\lambda_x(k)} - \frac{a_k^2}{\lambda_d(k)}\right) I_0\left(\frac{2a_k R_k}{\lambda_d(k)}\right) da_k}{\int_0^\infty a_k \exp\left(-\frac{R_k^2}{\lambda_d(k)}\right) \exp\left(-\frac{a_k^2}{\lambda_x(k)} - \frac{a_k^2}{\lambda_d(k)}\right) I_0\left(\frac{2a_k R_k}{\lambda_d(k)}\right) da_k}
\end{aligned} \tag{A.74}$$

Let us define,

$$\frac{1}{\lambda(k)} \triangleq \frac{1}{\lambda_x(k)} + \frac{1}{\lambda_d(k)} \tag{A.75}$$

$$v_k \triangleq \frac{\xi_k}{1 + \xi_k} \tag{A.76}$$

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \tag{A.77}$$

$$\gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)} \tag{A.78}$$

Thus, (A.74) may be simplified as,

$$\hat{A}_k = \frac{\int_0^\infty a_k^2 \exp\left(-\frac{a_k^2}{\lambda(k)}\right) I_0\left(2a_k \sqrt{\frac{v_k}{\lambda(k)}}\right) da_k}{\int_0^\infty a_k \exp\left(-\frac{a_k^2}{\lambda(k)}\right) I_0\left(2a_k \sqrt{\frac{v_k}{\lambda(k)}}\right) da_k} \tag{A.79}$$

To solve these integrals we resolve to the table of integrals provided in [37]. From (6.631.1) of [37], we have

$$\begin{aligned}
\int_0^\infty x^\mu e^{-\alpha x^2} J_\nu(\beta x) dx &= \frac{\beta^\nu \Gamma\left(\frac{1}{2}\nu + \frac{1}{2}\mu + \frac{1}{2}\right)}{2^{\nu+1} \alpha^{\frac{1}{2}(\mu+\nu+1)} \Gamma(\nu+1)} {}_1F_1\left(\frac{\nu+\mu+1}{2}; \nu+1; -\frac{\beta^2}{4\alpha}\right); \\
&= \frac{\Gamma\left(\frac{1}{2}\nu + \frac{1}{2}\mu + \frac{1}{2}\right)}{\beta \alpha^{\frac{1}{2}\mu} \Gamma(\nu+1)} \exp\left(-\frac{\beta^2}{8\alpha}\right) M_{\frac{1}{2}\mu, \frac{1}{2}\nu}\left(\frac{\beta^2}{4\alpha}\right) \tag{A.80} \\
&[\text{Re } \alpha > 0, \text{Re } (\alpha + \mu) > -1]
\end{aligned}$$

Where ${}_1F_1(\alpha; \gamma; z)$ is the confluent hypergeometric function. Now using the expression from (8.406.3) of [37] we have:

$$I_n(z) = i^{-n} J_n(iz) \tag{A.81}$$

$$\begin{aligned}
\int_0^{\infty} x^{\mu} e^{-\alpha x^2} I_{\nu}(\beta x) dx &= i^{-\nu} \int_0^{\infty} x^{\mu} e^{-\alpha x^2} J_{\nu}(i\beta x) dx; & (A.82) \\
&= i^{-\nu} \frac{(i\beta)^{\nu} \Gamma\left(\frac{1}{2}\nu + \frac{1}{2}\mu + \frac{1}{2}\right)}{2^{\nu+1} \alpha^{\frac{1}{2}(\mu+\nu+1)} \Gamma(\nu+1)} {}_1F_1\left(\frac{\nu+\mu+1}{2}; \nu+1; \frac{\beta^2}{4\alpha}\right); \\
&= \frac{\beta^{\nu} \Gamma\left(\frac{1}{2}\nu + \frac{1}{2}\mu + \frac{1}{2}\right)}{2^{\nu+1} \alpha^{\frac{1}{2}(\mu+\nu+1)} \Gamma(\nu+1)} {}_1F_1\left(\frac{\nu+\mu+1}{2}; \nu+1; \frac{\beta^2}{4\alpha}\right); \\
&\quad [\text{Re } \alpha > 0, \text{Re } (\alpha + \mu) > -1]
\end{aligned}$$

For $\nu = 0$ the expression reduces to

$$\begin{aligned}
\int_0^{\infty} x^{\mu} e^{-\alpha x^2} I_0(\beta x) dx &= \frac{\Gamma\left(\frac{\mu+1}{2}\right)}{2\alpha^{\frac{1}{2}(\mu+1)}} {}_1F_1\left(\frac{\mu+1}{2}; 1; \frac{\beta^2}{4\alpha}\right); \\
&\quad [\text{Re } \alpha > 0, \text{Re } (\alpha + \mu) > -1]
\end{aligned}$$

Now we solve the integrals of (A.79) substituting $x = a_k$, $\alpha = \frac{1}{\lambda(k)}$, $\beta = 2\sqrt{\frac{\nu_k}{\lambda(k)}}$ and the appropriate value for μ in (A.83).

$$\begin{aligned}
\hat{A}_k &= \frac{\frac{1}{2}\lambda(k)^{1.5} \Gamma(1.5) {}_1F_1(1.5; 1; \nu_k)}{\frac{1}{2}\lambda(k) {}_1F_1(1; 1; \nu_k)} \\
&= \frac{\Gamma(1.5) \lambda(k)^{1/2} {}_1F_1(1.5; 1; \nu_k)}{e^{-\nu_k}} \\
&= \lambda(k)^{1/2} \Gamma(1.5) {}_1F_1(-0.5; 1; -\nu_k) & (A.83)
\end{aligned}$$

Where we have used the substitutions:

$$\begin{aligned}
{}_1F_1(1; 1; z) &= e^z \\
{}_1F_1(\alpha; \gamma; z) &= e^z {}_1F_1(\gamma - \alpha; \gamma; -z) \quad [\text{From (9.212.1) of [37]}]
\end{aligned}$$

The MMSE estimator as expressed in the original paper of Empraim and Malah is given by:

$$\hat{A}_k = \Gamma(1.5) \exp\left(-\frac{\nu_k}{2}\right) \left[(1 + \nu_k) I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right) \right] R_k \quad (A.84)$$

