Faculty of Engineering & Science
Department of Electrical & Computer Engineering

Methods for Speech Intelligibility Enhancement

Maneesh Kumar Singh

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

October 2017

मेरे बडे पापा स्व. श्री राम नारायण सिंह “साधू” को समर्पित।

Dedicated to my uncle Late Sri Ram Narayan Singh "SAADHU".

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

The research presented and reported in this thesis was conducted in accordance with the National Health and Medical Research Council National Statement on Ethical Conduct in Human Research (2007) - updated March 2014. The proposed research study received human research ethics approval from the Curtin University Human Research Ethics Committee (EC00262), Approval Number **CSEA 300914**.

Date: October 9, 2017 Signature: .....

[Maneesh Kumar Singh]

Acknowledgments

"I am going to start my dissertation in respect of God,
Who is Gracious and Merciful."

Successfully completing any task gives us satisfaction as well as internal strength for future problems but the person alone has never existed. He is truly accompanied by few people. They use to give the personal support as well as suggestions to complete his work successfully. So I am pleasured in thanking all such people who motivate me and provides their kind support at all stages of my research work.

I was lucky enough to have a person cum mentor whose unique insights were a constant source of inspiration for my research work, and whose relaxed way of supervision allowed me to shape my research the way I wanted. This person is no other than my supervisor Dr. Siow Yong Low. Thank you so much Siow sir! I would like to extend my sincere gratitude to Dr. Zhuquan Zang for keeping a discrete eye on my research and offering me the assurance that I was going about it in a sensible way. I am also grateful to Prof. Sven Nordholm for enhancing my thinking ability, to a stimulating discussion with a very knowledgeable person in the field of speech processing.

Special thanks go to the people involved in my project, part of which was my PhD. I respectfully thank to Curtin Sarawak Research Institute (CSRI) director Prof. Clem Kuek, and my thesis chairperson Dr. Lenin Gopal for the throughout support.

Furthermore, the completion of my degree would have been a lot harder if it weren't for the help, support, and friendship of all the people I shared office with and during my stay. I particularly thanks, Prof. Ashutosh, Nikhil Raj, Arshad, and Bapi for making my life much more comfortable and a lot more interesting

life in Curtin.

I am thankful to the Information & Communication Technology (ICT), Curtin University, particularly, Megawati and her team for providing me an excellent support and friendly environment for all my simulation works. Without their valuable support, it would not be possible to conduct my long-run computer-based simulations.

I would like to thank my brother in law Mr. Manish Singh for being supportive throughout, and helping me in all the possible ways. A special thanks to my wife Maneesha Singh and my sweet daughter Manvisha Singh Chauhan (Chitu), the words cannot express how grateful I am, for all of the sacrifices that you both have made on my behalf. Your prayer was what sustained me thus far.

Finally, if I have forgotten anyone, I apologize.

Maneesh K. Singh

October 9, 2017

$\sim * \sim$

Abstract

The exchange of information via speech is nowadays possible from almost all places at any time. However, even though the vision of permanent reachability and connectivity has been realized in the meantime nearly worldwide, there is still room for improvements when it comes to the transmission of speech under noisy conditions. The performance of any speech communication system may significantly deteriorate when the speech signal is disturbed by ambient interferences such as traffic noise or office noise, possibly leading to a reduced speech quality and intelligibility. Recently, modulation domain has been reported to be a better alternative to the time-frequency (acoustic) domain for speech enhancement, as speech intelligibility is highly correlated with the modulation spectrum even in noisy conditions. This suggests modulation spectrum may assist in the demarcation of speech and noise. Motivated by that, this thesis investigates the use of modulation domain for estimating the noise spectral amplitudes, which consists of three main parts.

In the first part, we acknowledge the fact that the Gaussian assumption for all noise DFT coefficients does not necessarily hold, and therefore, the best noise distribution which will be suitable for both the acoustic and modulation domain based speech applications has been investigated. Results show that the modulation based Gamma density function better represents the noise density for both stationary and non-stationary noise signals compared to the non-modulation domain. The modulation based Gamma density is then used to derive the noise estimator via Bayesian motivated MMSE approach. As the modulation based estimation closely matches the true density of the noise, the proposed noise estimator does not require bias compensation even for poor signal-to-noise ratio (SNR) conditions, i.e., $\leq$ 5 dB. The proposed Gamma based noise estimator

achieves higher noise suppression against conventional methods in terms of perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and segmental SNR (SNRseg) measures of the enhanced speech.

In the second part of this thesis, we are concerned with the extension of the Bayesian estimators to estimate the noise DFT coefficients and to explore their applicability in the short-time modulation domain. Frankly speaking, the subjective meaningful Bayesian methods are available only for the speech estimation, as the noise estimators based on perceptually motivated Bayesian cost functions are still elusive. Therefore, we consider the derivation of the family of estimators by considering the perceptual aspect of the Bayesian cost functions, which provides the better tracking of the time-varying noise signals in the short-time modulation domain. The main outcome of the theoretical framework is a noise estimator that exploits some similarities with the parent speech estimator.

In the final part of this thesis, the derived Bayesian estimators have been implemented for estimating the noise signals with the aim of, firstly, gaining a better understanding of their properties towards modulation domain, and secondly, the role of the parameters used in an analytical generalization of the estimator's structure. These chosen parameters based on the characteristics of the human auditory system are found to have a good correlation in the tracking of the non-stationary noise by providing the better performance while limiting the speech distortions at low input SNR levels.

$\sim * \sim$

Contents

List of Figures

List of Tables

List of Symbols

General Symbols

E[.]	expectation operator
$\hat{(.)}$	estimated value
P[.]	probability density function
$\Re(.)$	real-part of a complex variable
$Im(.)$	imaginary part of a complex variable
$\angle(.)$	angle of a complex variable
log(.)	natural logarithm
min[.]	minimum value
$e^{(.)}$	exponential function exp(.)
$\approx$	approximately
$\lvert . \rvert$	absolute value
$\sum$	summation of

$\int$	integration of
$y(t)$	noisy speech in discrete-time domain
$x(t)$	clean speech in discrete-time domain
$d(t)$	noise in discrete-time domain
D_e	binary decision of the VAD
E_t	energy threshold value
E_r	energy of the most recent unvoiced frame
α	recursive smoothing parameter
α_c	smoothing constant
Λ	probability of noise-only frame
ξ_{H_1}	fixed a priori SNR
$\gamma_{ig}(\mathsf{x}, \mathsf{a})$	incomplete gamma function of x
$\mathcal{P}$	exponential time-frequency smoothing constant
λ_g	Gaussian noise variance
λ_r	Rayleigh noise variance
ν_g	shape parameter in Gamma distribution
β_g	scale parameter in Gamma distribution
$\Gamma(.)$	Gamma function
$f_{N\|Z}(n\|z)$	a posterior noise probability density function
$f_N(n)$	noise probability density function
$\mathcal{R}(n, z)$	Bayesian risk function

$C(\hat{n}, n)$	Bayesian cost function
$I_0(.)$	modified Bessel function of order zero
$I_1(.)$	modified Bessel function of order one
$D_{\mathsf{n}}(\mathsf{x})$	parabolic cylinder function of order n
σ^2_{mn}	noise expectation estimate in modulation domain
$\mathcal{E}_n$	noise estimation error
$\Phi(a, b, \mathsf{x})$	confluent hypergeometric function
$\forall$	for all
γ_e	Euler's constant
$\zeta(.)$	Weierstrass's zeta function
$\frac{\delta}{\delta \mathsf{x}}(.)$	partial differentiation

Symbols in Frequency Domain

$\alpha_d(l, k)$	time-varying smoothing constant
$\alpha_{opt}(l, k)$	optimum time-varying smoothing constant
$S_{pp}(l, k)$	speech presence probability
$\xi(l, k)$	a priori SNR
$\gamma(l, k)$	a posteriori SNR
$S_{sig}(l, k)$	sigmoid function
B_c(l,k)	biasing compensation factor

l	time frame index in frequency domain
k	acoustic frequency bin index
K	acoustic frame duration
S	acoustic frame shift (%)
$\mathbf{Y}(l,k)$	complex-valued DFT coefficients of noisy speech in frequency domain
$Y(l,k)$	noisy speech magnitude coefficients in frequency domain
$\lambda_{ay}(l,k)$	noisy variance in frequency domain
$\mathbf{X}(l,k)$	complex-valued DFT coefficients of noise in frequency domain
$X(l,k)$	clean speech magnitude coefficients in frequency domain
$\lambda_{ax}(l,k)$	clean speech variance in frequency domain
$\mathbf{D}(l,k)$	complex-valued DFT coefficients of clean speech in frequency domain
$D(l,k)$	noise magnitude coefficients in frequency domain
$\lambda_{ad}(l,k)$	noise variance in frequency domain
$w_{aa}(t)$	analysis window in frequency domain
$w_{as}(t)$	synthesis window in frequency domain

Symbols in Modulation Domain

τ	modulation time frame index
m	modulation frequency index
M	modulation frame duration w.r.t. acoustic frame duration

P	modulation frame shift (%)
$\lambda_{md}(\tau, k, m)$	variance of noise in modulation domain
$w_{ma}(t)$	analysis window in modulation transform
$w_{ms}(t)$	synthesis window in modulation transform
$\mathbf{Z}(\tau, k, m)$	complex-valued DFT coefficients of noisy speech in modulation domain
$Z(\tau, k, m)$	noisy speech magnitude coefficients in modulation domain
$\lambda_{mz}(\tau, k, m)$	variance of noisy speech in modulation domain
$\mathcal{X}(\tau, k, m)$	complex-valued DFT coefficients of clean speech in modulation domain
$\mathcal{X}(\tau, k, m)$	clean speech magnitude coefficients in modulation domain
$\lambda_{mx}(\tau, k, m)$	variance of clean speech in modulation domain
$\mathbf{N}(\tau, k, m)$	complex-valued DFT coefficients of noise in modulation domain
$N(\tau, k, m)$	noise magnitude coefficients in modulation domain
$\theta_{mz}(\tau, k, m)$	modulation phase spectrum of noisy speech
$\theta_{mx}(\tau, k, m)$	modulation phase spectrum of clean speech
$\theta_{mn}(\tau, k, m)$	modulation phase spectrum of noise

List of Acronyms

AFFT	acoustic fast Fourier transform
AFS	acoustic frame shift (%)
AM	amplitude-modulation
AMS	analysis-modification-synthesis
CDF	cumulative distribution function
dB	decibel
DD	decision-directed
DFT	discrete Fourier transform
DSP	digital signal processing
EM	expectation-maximization
FFT	fast Fourier transform
GOF	goodness-of-fit
IDFT	inverse DFT
IMCRA	improved minima controlled recursive averaging
IS	Itakura-Saito
ISTFT	inverse short-time Fourier transform

kHz	kilohertz
KS	Kolmogorov-Smirnov
KurtR	kurtosis ratio
LED	linear energy-based detector
LLR	log-likelihood ratio
LMS	least mean square
LogErr	symmetric logarithmic-error distortion measure
LSA	log spectral amplitude
MAP	maximum a posteriori
MCRA	minima controlled recursive averaging
MFS	modulation frame shift (%)
MFFT	modulation fast Fourier transform
ML	maximum likelihood
MMSE	minimum mean square error
MMSE-BC	MMSE with bias compensation
MMSE-UB	MMSE with unbiased compensation
MS	minimum statistics
MSE	mean square error
PDF	probability density function
PESQ	perceptual evaluation of speech quality
PSD	power spectral density

ROF	rate of fall
ROR	rate of rise
SIG	sigmoid
SMT	spectral minima tracking
SNR	signal-to-noise ratio
SNRseg	segmental SNR
SPP	speech presence probability
SPU	speech presence uncertainty
SS	spectral subtraction
STFT	short-time Fourier transform
STOI	short-time objective intelligibility
STSA	short-time spectral amplitude
TMTF	temporal modulation transfer function
VAD	voice activity detection
WCOSH	weighted COSH
WE	weighted Euclidean
WF	wiener filter
WGN	white Gaussian noise
ZCR	zero-crossing rate
β-MMSE	β-order MMSE

This Page Intentionally Left Blank.

Chapter 1

Introduction

The greatest musical instrument given to a man is the voice.

–Dayananda Saraswati.

1.1 Introduction

The transfer of information through speech communication has been made popular by the use of speech-processing based devices like cellular phones [1–3], digital hearing aids [4–10] and various human-to-machine speech processing applications [11–15]. With the increased use of these speech communication devices, there is a strong need for improvements when it comes to the transmission of speech under noisy conditions. This is because noise is everywhere and the performance of any speech communication system significantly deteriorates when a noisy location such as factory, restaurants and other places of social gathering [16].

The most common sources of noise are the additive background noise, which is always present in different degrees in any location. For example, operating a hands-free mobile phone in a car can be affected by at least three types of background noise, namely wind, road as well as engine noise. Other examples of noisy speech inputs are such as food courts and bus terminals, speech communication systems in cockpits, cellular phones in a factory, which, therefore, degrades the resulting speech at the receiver end. As a result, these speech-based devices are potentially exposed because, a common complaint among users is the inability

to focus on a single speaker, especially in situations with multiple interfering speakers. Due to these interfering noise signals, the speech characteristics are modified, and the effect will profoundly affect the listener's perception of the observed speech. Moreover, due to the complexity and highly non-stationary nature of the speech signals, estimation of the speech DFT coefficients has been still a challenging problem [17–21].

Therefore, to make speech communication possible, natural, and comfortable under noisy conditions, it is highly desirable to develop methods to mitigate these background noise effects and restore the original speech successfully. This problem of reducing the noise signals from the noisy speech is referred to as either speech enhancement or noise reduction.

Generally speaking, the methods of enhancing the speech signal are divided into two major categories, single microphone, and multi-microphone based noise reduction methods. In the single-microphone methods, the realization of speech and noise signals are obtained by using the single microphone. The fundamental problem with these methods is to achieve noise reduction by analyzing and processing the noisy speech measured by only one microphone without requiring any other additional information. Since both speech and noise signals are filtered at the same time, demarcation process is the most critical, yet most challenging issue in the field of speech enhancement. However, considering cost and size, these methods are very useful nowadays, especially in mobile communication, where only one microphone is available.

The multi-microphone methods on the other hand utilize more than one microphone and achieve better performance. Although these methods often lead to better performance than single-microphone methods, the usability is limited by additional costs of the microphone, power usage, computational complexity, and size demands which are not always possible to implement on small devices.

Since speech is a highly non-stationary signal, the time domain realization is only possible on a short-time basis, typically in the order of few tens of milliseconds so that the signal is stationary for each time-frame. However, the original problem of handling the non-stationary speech signal in time domain is circumvented by converting the signal into the frequency domain by segmenting the noisy

signal into a short-time frame and by taking the Fourier transform (STFT). The respective Fourier coefficients of the noisy speech are relatively slow varying and therefore, it is easy to perform any spectral modification by providing an appropriate weighting (gain) function. The advantage of performing the noise estimation process in the frequency domain is that the structure of speech enhancement method in frequency domain handles different frequencies independently, that allows an appealing flexibility to exploit the noise statistics and improves the quality and intelligibility of the noisy speech [17, 19, 22].

Beyond the context of the frequency domain, several approaches to solve the single-microphone (single-channel) problems have been developed in the short-time modulation frequency domain [23–27]. Schimmel in his dissertation [28] reports that the energy from two different signals (e.g., speech and noise signals) in the modulation domain is largely non-overlapping. This is also supported by psychoacoustic research, which indicates that the human auditorial system segregates sound in the modulation domain [29]. This suggests greater demarcation between the noise and the speech in the modulation domain. Also, various research, which dates back to the early 90s, indicate that modulation domain processing often results in higher intelligibility of speech [27, 30–34]. Moreover, the intelligible components of the speech signal are mostly confined to the modulation frequency band of 1Hz to 16Hz and therefore processing can be made to concentrate on the relevant bands [35, 36].

Although modulation domain helps in improving the overall intelligibility of speech and holds a great potential especially for the single-microphone speech enhancement based applications [27, 34, 37], selection of appropriate frame length remains a topic that is the subject of ongoing research. The reason may be that the selection of frame length (FFT size) for both frequency and modulation domains based speech enhancement depends on the application.

Secondly, the key assumption in noise estimation is that the noise spectral coefficients are assumed to follow the Gaussian distribution [20, 38, 39]. Whilst the Gaussian assumption may be sufficient for stationary noise, it may not be the case for non-stationary noise signals such as babble or heavy street noise. This is because the Gaussian assumption for the speech spectral coefficients holds asymp-

totically only for long duration analysis frames [16]. Therefore, owing to speech non-stationarity, speech spectrum coefficients need to be estimated by using a shorter window frame (e.g., $20 - 40$ ms) to reflect the new statistics. Also, research shows that non-Gaussian functions such as Laplacian or Gamma are more accurate models for speech spectral coefficients [40–42]. In a similar way, environmental noise such as restaurant or street noise is time-varying and thus the characterization of noise density by using the Gaussian assumption may not be adequate. This complication can be reduced by using the Bayesian estimation theory [20, 21]. This is because the optimal estimators can be obtained by minimizing the Baye's risk function, which includes a posterior probability model of the unknown parameters (given from the observation vector) and a cost error function. The posterior probability density function (pdf) depends on how relatively the noise pdf is peaked, i.e., the likelihood pdf depends on the posterior pdf. Generally, the more peaked the noise pdf, the larger the estimation error will be, and as a result, the greater the influence on the outcome of the noise estimation process [43].

Therefore, this thesis utilizes the subjective meaningful Bayesian methods for noise estimation. This motivation is from the fact that these Bayesian methods are available only for the speech spectral estimation. In other words, the noise estimators based on perceptually motivated Bayesian cost functions are still elusive. Therefore, to explore the role of Bayesian based noise methods for all time-varying noise signals, perceptually motivated family of noise estimators have been derived in this thesis.

1.2 Objective

The scope of the Bayesian motivated noise estimation explored in this thesis is mainly focused on the tracking of the background noise by using the following main objectives:

- to investigate the use of modulation domain for noise modeling and estimation compared to the frequency domain,
- to provide the most appropriate noise density function for all time-varying

noise signals, which is remain inconclusive as no single noise density function can represent the different real world noise correctly, and

- to develop the adaptive modulation based Bayesian noise estimators for tracking of noise statistics applicable to all-time varying noise sources in the modulation domain.

Considering the feasibility and easy implementation of modulation domain based noise estimators to the speech processed devices, this thesis uses only one microphone in order to minimize the cost, size and power usage. To make a more robust and adaptive design of speech based devices, the proposed modulation domain noise methods in this thesis do not use

- knowledge of the speech signal,
- assumptions for the speech DFT coefficients and
- additional method to detect speech active or inactive periods (no-VAD).

1.3 Thesis Contribution

By achieving the above objectives, we extend the knowledge of modulation domain for time-varying noise estimation by using the proposed noise estimator and several new Bayesian extended noise estimators, that show the advantage over existing noise methods in the modulation domain. The main contributions of this thesis work are summarized below.

1.3.1 Modulation Frame Length Selection

The suitability of modulation frame length (FFT size) towards speech intelligibility is still twofold because different speech based applications use different modulation FFT size. For example, hearing aids devices requires intelligibility of the speech whilst, transmitting the signal through a communication channel requires better speech quality. Besides that, smaller modulation FFT size (MFFT) ($\leq$ 128) provides better speech intelligibility whilst, larger MFFT size ($\geq$ 256) introduces spectral roughness (smearing), which is audible as a distortion. In

this thesis, various combinations of both frequency and modulation FFT size and frame-shift have been employed in order to achieve the suitable modulation framework for single-channel speech enhancement system. Results from numerous investigative experiments reveal that by using modulation FFT size 32 and 64 with frame shift 50% achieves higher speech intelligibility. Whilst increasing the modulation FFT size lowers the intelligibility of the enhanced speech.

1.3.2 Noise Spectral Density Function

One key assumption in noise estimation is that the noise spectral coefficients are assumed to follow the Gaussian distribution. Whilst, the Gaussian assumption may be sufficient for stationary noise, it may not be the case for time-varying real world noise signals, such as interfering talkers originated in a restaurant or in a social gathering. Although the Gaussian assumption facilitates a mathematically tractable derivation for noise spectral estimators, these time-varying noise signals can not be Gaussian distributed, and therefore, the characterization of the time-varying probability distribution of noise is still inconclusive, as no single noise density function can represent the different real world noise in the DFT domain. From the experiments, it is noticed that the Gamma density models all time-varying noise signals more appropriately that results better noise estimates compared to the state of the art methods.

1.3.3 Family of Bayesian Noise Estimators

Interestingly, for the speech spectral estimation, the perceptually meaningful Bayesian cost functions have been utilized by arguing that these cost functions correlate to the human auditory system. From spectral stationarity point of view, if the speech DFT coefficients can be estimated with these cost functions, there would be a significant attention to perform the noise estimation by using the same Bayesian cost functions. Therefore, these perceptually meaningful Bayesian cost functions are extended for deriving the family of noise estimators. Since the noise DFT coefficients are comparably more stationary than the speech DFT coefficients, tracking the noise DFT coefficient may be easier by using family of noise estimators as these generalized noise estimators give the flexibility to

choose appropriate parameter to achieve the optimum performance in terms of the speech intelligibility.

1.3.4 Importance of Modulation Domain

The modulation framework selection for improving the speech quality and intelligibility is motivated by the fact that, different modulation based applications use different FFT sizes. As smaller FFT size provides higher intelligibility, therefore speech based applications such as hearing aid devices prefer smaller FFT size. Applications such as speech coding, on the other hand, prefer better quality over intelligibility and, the selection of modulation frame length may differ by having a larger FFT size. In a similar way, the performance of the noise estimator in the modulation domain may differ from estimator to estimator. In this thesis it is found that the proposed noise estimators in modulation domain adapt the noise spectral changes efficiently without requiring any bias compensation factor. This is due to fact that the modulation domain provides slow spectral variation compared to frequency domain. The spectral stationarity test conducted in this thesis, clearly, reveals that the noisy speech spectrum in the modulation domain is slow varying as compared to the spectral variation in the frequency domain. This slow spectral variation allows sufficient time to the estimator to adapt any spectral changes that helps to improve the speech intelligibility.

1.4 Thesis Outline

The structure of this dissertation is as follows.

In Chapter 2, we review the most prominent approaches related to single-channel speech enhancement. Naturally, we focus our attention on the different noise estimation methods of the frequency domain. We also address topics related to modulation based speech enhancement and, important aspects of modulation domain towards noise estimation by using a single-microphone. The frameworks of both frequency and modulation domain based single-channel speech enhancement methods are also included in this chapter.

Chapter 3 formulates the problem of enhancing the noisy speech as the time-

varying noise density assumption problem. After fitting the distributions to both frequency and modulation DFT coefficients of all types of noise signals (stationary and non-stationary), we derive the proposed noise estimator that provides the better noise estimate with improved speech quality and intelligibility in both frequency and modulation domains. This chapter apart from providing a noise estimation methods that improves the segmental signal-to-noise ratio (SNR_{seg}) and speech intelligibility, also focuses on the selection of appropriate FFT size and shift that register the modulation domain for improved speech intelligibility.

In Chapter 4, the perceptually meaningful Bayesian cost functions are extended for deriving the family of estimators for estimating the noise spectral magnitudes. These generalized estimators represent the family of estimators and therefore provide more flexibility to choose the appropriate parameter to achieve the optimum performance in terms of the speech intelligibility.

Chapter 5 presents the experimental performance of the Bayesian motivated noise estimators derived in chapter 4. Particularly, we were interested in identifying and studying the behavior of these Bayesian noise estimators in comparison with the existing parent speech estimators in the sort-time modulation domain. It was motivated in order to find the suitable values of the parameters such as β and the exponent p used in these noise estimators for modulation domain based speech enhancement.

Finally, chapter 6 summaries the work presented in this thesis and draw the conclusions that have stemmed from this work.

1.5 Experimental Considerations

1.5.1 Speech Source

The experimental setup in this thesis employs a set of 12 phonetically-balanced clean speech sentences from the TIMIT database [44]. This set of clean speech sentences belongs to six male speakers and six female speakers. The selected male and female speakers are chosen differently, which are originally sampled at 44.1kHz. To simulate the receiving frequency characteristics of telephone handsets, the corpus is then down-sampled to 8 kHz. Moreover, the length of each

sentence is 4.5s, whilst 0.5s of initial silence is added further to make sure that the initial noise-only frame is long enough for noise statistics.

1.5.2 Noise Source

Considering the noisy situations happening nowadays in our real life, four different noise sources with different time-varying stationarity namely; stationary white noise, long term stationary factory noise, heavy street noise, and highly non-stationary babble noise signals are considered from NOISEX-92 database [45], in order to cover all practically originated real world noise signals. These all time-varying noise signals are added to the clean speech at a wide range of input SNRs, i.e., 0, 5, 10, 15 and 20dB.

1.5.3 Speech Quality and Intelligibility Measures

Objective Measures

Many objective measurement algorithms have been derived in the literature for evaluating the performance of speech enhancement algorithms [46,47]. The most widely used methods include the PESQ measure [48], and the SNRseg measure [49]. The PESQ measure, which was not originally designed to evaluate the performance of speech enhancement algorithms, has been found to have a good correlation overall with mean opinion score(MOS) [47]. It predicts the MOS scores which yields a result from 1 to 5, where a higher score indicates a better speech quality. Meanwhile, the SNRseg measure is also preferred among the vast amount of objective measures as it has been found to correlate best with background noise reduction [47].

Throughout this Thesis, both the PESQ measure and the SNRseg measure were used to evaluate the performance of the proposed algorithms. The PESQ measure was implemented based on the procedures presented in [46] whilst, the SNRseg measure is defined as [46]

$$\text{SNRseg} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\frac{||\mathrm{x}(m)||^2}{||\mathrm{x}(m)|| - ||\hat{\mathrm{x}}(m)||} \tag{1.1}$$

where the vector $\mathrm{x}(m)$ represents a clean speech (time-domain) frame, and $\hat{\mathrm{x}}(m)$ is the enhanced speech frame. In order to discard non-speech frames, each frame was threshold by a -10dB lower bound and a 35dB upper bound.

The performance of the speech enhancement scheme has a trade-off between musical noise, speech distortion and noise reduction. The PESQ measure and the SNRseg measure can not represent the whole picture of these trade-offs. Therefore, an objective Short-Time Objective Intelligibility (STOI) measure [50] is also utilized to evaluate and compare the results between the amount of musical noise, speech distortion and noise reduction generated from the speech enhancement scheme.

1.6 Publication

The following paper is accepted in *Speech Communication* in conjunction with Chapter 3 of this thesis.

Maneesh K. Singh, S. Y. Low, S. Nordholm and Zhuquan Zang, *"Bayesian Noise Estimation in the Modulation Domain"*, Speech Communication. [*Accepted*].

~ * ~

Chapter 2

Literature Review

Arise! Awake! and don't stop until your goal is achieved.

–Swami Vivekananda.

2.1 Introduction

In the past decades, various techniques have been developed for single-channel noisy speech enhancement. These algorithms can be grouped according to the theory on which they are based into the categories such as estimating either the speech signal or the noise to improve the quality and intelligibility of the noisy signal. The methods for estimating speech signals can be grouped as spectral-subtractive methods [18, 19], statistical-model-based methods [20, 38, 40, 51], subspace methods [52, 53], Kalman filtering methods [54, 55], etc. Whilst, the noise estimation methods can be categorized by voice activity detection (VAD) [56–58], and the non-VAD based methods such as the spectral minima tracking [59], minimum statistics (MS) [60, 61], time-recursive averaging [62–64], and statistical-model-based noise estimation methods [39, 65–67].

This chapter presents an overview of noise estimation methods while maintaining a focus on those, which are the non-VAD based noise estimation methods in the frequency domain, as those are the methods that are central to this dissertation.

2.2 Signal Model

For a single-channel based speech enhancement, many popular methods employ the analysis-modification-synthesis (AMS) framework to perform the speech enhancement in frequency domain [46,68–70]. The AMS framework consists of three stages, namely analysis, modification, and synthesis stage. The analysis stage of the framework segments the noisy speech into the short-time frames and the magnitude and phase spectral components is achieved by using the short-time Fourier transform (STFT). The magnitude or phase spectrum is used for the modification to improve the speech quality and intelligibility in the modification stage and, finally, the inverse STFT followed by the overlap-add synthesis is used to reconstruct the enhanced speech [68, 69, 71] to the time domain. The block diagram representation of a single channel speech enhancement framework is shown in Fig. 2.1 and the simplified AMS based single-channel speech enhancement framework is shown in Fig. 2.2.

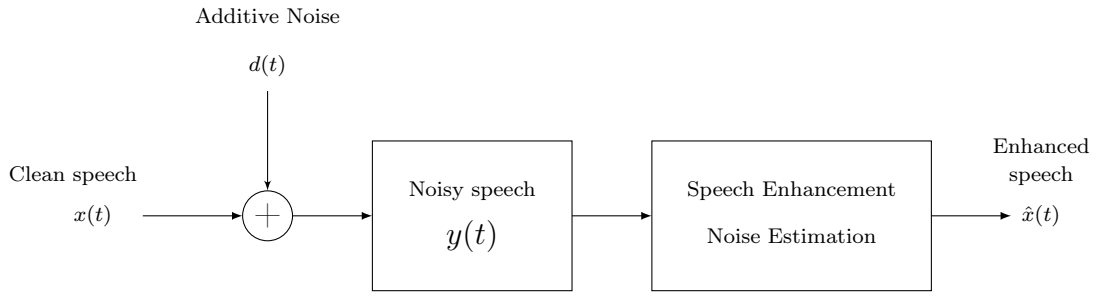


Figure 2.1: The block diagram representation of a single-channel speech enhancement system.

Let the noisy speech signal $y(t)$ in the time domain be denoted as

$$y(t) = x(t) \; + \; d(t), \tag{2.1}$$

where, $x(t)$ is the clean speech signal corrupted by uncorrelated additive noise $d(t)$, which is assumed to be a zero-mean. The short-time frequency representation of $y(t)$ is therefore given by

$$Y(l,k) = \sum_{t=0}^{K-1} y(t+lS) w_{aa}(t) \; e^{-j\frac{2\pi tk}{K}}, \tag{2.2}$$

where $0 \leq k \leq K$, k corresponds to the acoustic frequency bin index, K is the

acoustic FFT size, S denotes the acoustic frame shift in samples between successive window frames (%) and, $w_{aa}(t)$ is the acoustic analysis window. Since the frames are overlapping, the respective percentage (%) of the current frame samples have to be added with the previous frame, which is known as the overlap-add method. The benefit of overlapping frames is that a smooth transition between consecutive frames can be achieved [27, 34, 71]. Since, noise is assumed to be uncorrelated and additive, applying the linearity property of DFT coefficients to frequency domain spectrum of noisy speech yields

$$Y(l,k) = X(l,k) + D(l,k). \tag{2.3}$$

Similarly, the respective magnitude spectrum of the noisy speech can be represented as

$$|Y(l,k)| = |X(l,k)| + |D(l,k)|, \tag{2.4}$$

where, $Y(l,k)$, $X(l,k)$, and $D(l,k)$ represent the complex-valued DFT coefficients of the noisy speech, clean speech and additive noise, respectively, and $|Y(l,k)|$, $|X(l,k)|$, and $|D(l,k)|$ are their respective magnitude spectrum. Since both speech and noise signals are uncorrelated by assumption, the variance of $|Y(l,k)|$ is

$$\lambda_{ay}^2(l,k) = E\big[|Y^2(l,k)|\big] \tag{2.5}$$

$$= \lambda_{ax}^2(l,k) + \lambda_{ad}^2(l,k), \tag{2.6}$$

where $E\big[.\big]$ denotes the mathematical expectation, and

$$\lambda_{ax}^2(l,k) = E\big[|X^2(l,k)|\big], \tag{2.7}$$

$$\lambda_{ad}^2(l,k) = E\big[|D^2(l,k)|\big], \tag{2.8}$$

are the variances of $|X(l,k)|$ and $|D(l,k)|$, respectively.

After the modification stage of the AMS framework, the modified complex-valued DFT coefficients of speech spectrum $\hat{X}(l,k)$, can then be found by combin-

ing the modified magnitude $|\hat{X}(l,k)|$ of speech and noisy phase spectra $\angle Y(l,k)$ as

$$\hat{X}(l,k) = |\hat{X}(l,k)|e^{j\angle Y(l,k)}, \tag{2.9}$$

and, the enhanced speech is synthesized by applying inverse STFT to $\hat{X}(l,k)$, followed by overlap-add synthesis [70], as

$$\hat{x}(t) = \sum_{l} \left\{ w_{as}(t - lS) \sum_{k=0}^{K-1} \hat{X}(l,k)\ e^{j\frac{2\pi(t-lS)k}{K}} \right\}, \tag{2.10}$$

where, $w_{as}(t)$ is the synthesis window function. The window functions $[w_{aa}(t) = w_{as}(t) = w(t)]$ have to be chosen such that the perfect reconstruction of enhanced speech is achieved.

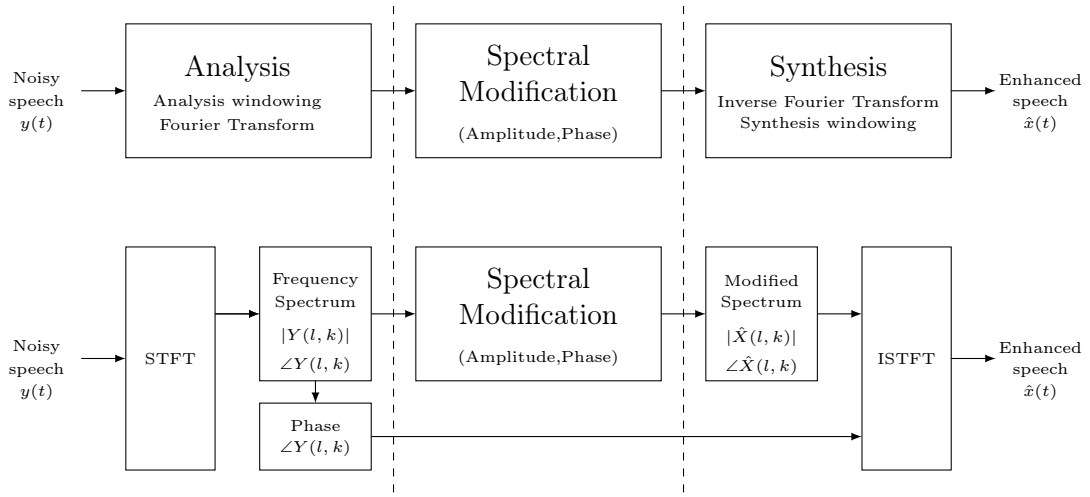


Figure 2.2: The simplified block diagram representation of an AMS framework for frequency domain speech enhancement.

The existing AMS framework of speech enhancement algorithms only provides the modification of the magnitude spectrum while keeping the phase spectrum unchanged. The role of phase spectrum in speech enhancement is twofold, where the assumption of keeping the phase spectrum unchanged is based on a long-standing belief that for small window durations, typically 20ms to 40ms, the short-time phase spectrum carry useful small information and is not important in speech enhancement [72–76]. On the other hand, recent studies claim that the phase estimation may further improve the limits of single-channel speech enhancement and results in higher intelligibility of enhanced speech [77–79]. By using the Hamming window function in short-time Fourier analysis, the magnitude spec-

trum contributes significantly more towards speech intelligibility compared to the phase spectrum [76, 80]. Therefore, we have employed the Hamming window for both the analysis and synthesis window functions throughout in this thesis to reduce the effect of the unchanged phase spectrum on speech intelligibility.

2.3 Noise Estimation methods

Noise estimates have a major impact on the quality of enhanced signal in speech enhancement applications. For instance, if the noise is under-estimated, an annoying residual noise will be audible, whilst over-estimating the noise will distort the original speech [81]. Generally speaking, noise estimation methods can be categorized into voice activity detector (VAD) based and non-VAD based noise estimation methods. In the non-VAD based methods, the best-known methods are probably the spectral minima tracking [59], minimum statistics (MS) [60], Minima-Controlled Recursive Averaging [63], the statistical-model-based noise estimation [39, 65–67]. These methods could indeed provide good alternatives to the VAD-based noise methods, as the noise spectrum can be continuously updated and thus better tracked, compared to only updating during noise-only periods.

2.3.1 Voice Activity Detection (VAD)

The basic principle of a voice activity detector (VAD) is to detect speech active periods with other period labeled as noise only. The VAD makes a binary decision on a frame-by-frame basis to detect speech active periods and, the noise estimate is obtained by recursively averaging noise during the speech pauses [56–58]. Here, the spectrum of the noise signal is assumed to be stationary between speech pause and processing periods, where the estimated noise is used. Such noise estimation methods strongly depend on the accuracy of the VAD. However, the performance of the VAD may not be reliable for poor signal-to-noise ratio (i.e., $\leq$5dB). This is particularly the case for non-stationary noise where a sudden rise in the noise power may be misinterpreted as a speech signal. The following sub-sections describe the process of a traditional voice activity detectors.

Energy Thresholding

In the energy-based VAD, the signal energy is compared with the threshold depending on the noise level and, the speech presence is detected when the estimated noisy frame energy E_y is higher than the threshold, as

$$D_e = \begin{cases} \text{noise only} & \text{if} \quad \text{if } E_y < \Lambda E_t \\ \text{speech+noise} & \text{if } E_y \geq \Lambda E_t. \end{cases} \tag{2.11}$$

The D_e is the binary decision of the VAD, represents 0 for a noise-only frame, whilst, 1 for a frame containing both speech and noise signals. The $\Lambda > 1$, is a scaling factor allows a safe band for adapting the threshold energy, E_t. Different energy-based VADs provides different methods to update the thresholds. The simplest energy-based method is the Linear Energy-Based Detector (LED) [82], where the energy threshold value is updated recursively, as

$$E_t(l,k) = (1-p)E_t(l-1,k) + p.E_r, \tag{2.12}$$

where, p is the smoothing parameter and, $E_t(l,k)$ is the current updated value of the threshold, whilst, E_r is the energy of the most recent unvoiced frame.

Zero Crossing Rate (ZCR)

In a zero crossing rate (ZCR), the VAD decision aims to calculate the number of times the signal amplitude crosses the x-axis in a given frame while considering the reference amplitude zero [83]. Since noise-only frame has less information than the noisy speech frames (as only noise information is available), the ZCR has a lower average value, and therefore, the decision is made that, if the ZCR for a given frame is below a certain threshold value, δ, it is assumed to be a noise-only frame. Otherwise, that frame contains speech as well as noise.

In [83], the VAD algorithm uses the ZCR along with the short-term energy of the noisy signal to detect the speech presence (or absence) for each frame. If the noisy signal energy $E_y(l)$ for a given frame l, rises above the average estimated noise energy $E_n(l)$, then it is likely that frame contains speech plus noise.

Otherwise, it is a noise-only frame, as follows

$$D_e = \begin{cases} 1 & \text{if } ZCR(l) > \delta \text{ and } E_y(l) > E_n(l) \\ 0 & \text{else.} \end{cases} \tag{2.13}$$

However, there are difficulties common to all VAD based methods [60]. Firstly, there is the situation where the noisy signal contains longer speech segment with very few speech pauses, that means to have the limited noise updates. In that time, the VAD based noise estimate may have varied sufficiently which turns into an inaccurate estimation. This results in a wrong noise estimation, which produces musical noise and distortion in the enhanced speech. Even theoretical VADs which perfectly decide between noise and speech frames can produce poor results if the speech pauses are too long or if the noise is fast varying.

Besides that, mostly the VADs have difficulty in differentiating noise and speech correctly at poor SNR conditions. This results in the estimated noise spectrum containing speech components which may attenuate the speech signal incorrectly, and suffer by the loss of speech information.

The non-VAD based methods are those which do not require detection of speech or noise frames. Often they will have an adaptive update parameter which controls the level of noise update when speech is present or absent. Therefore the noise estimate is continually updated throughout the signal, and not limited to regions where no speech is present, allowing much more frequent updating.

2.3.2 Spectral Minima Tracking

Doblinger [59], the first who developed a noise estimation technique by tracking the continuous minima of the noise spectrum without applying any VAD. The noise estimate in this method was updated continuously by smoothing noisy speech power spectra in each frequency bin separately by introducing a non-linear smoothing factor. In search of the noisy speech power spectrum minimum, a short-time smoothed version of the noisy speech power is estimated recursively by

$$\hat{\lambda}^2_{ay}(l,k) = \alpha\hat{\lambda}^2_{ay}(l-1,k) + (1-\alpha)|Y^2(l,k)|, \tag{2.14}$$

with a smoothing parameter α. The noise power estimate by using the non-linear

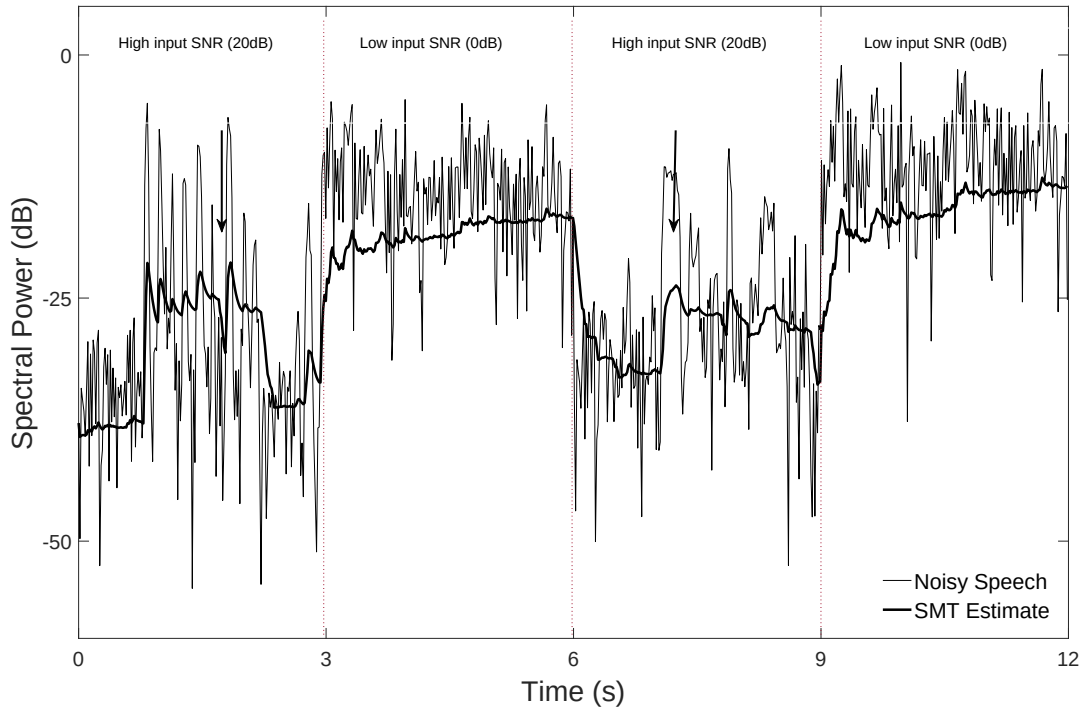


Figure 2.3: Plot of the noisy speech power with Spectral Minima Tracking (MST) estimate by using two input SNR conditions (0dB & 20dB) of babble noise at a given frequency 250Hz.

smoothing parameter is given after tracking the minimum of noisy speech power in each frequency bin separately, as follows

$$|\hat{D}^2(l,k)| = \begin{cases} \gamma|\hat{D}^2(l-1,k)| + \frac{1-\gamma}{1-\beta}\left[\hat{\lambda}^2_{ay}(l,k) - \beta\hat{\lambda}^2_{ay}(l,k)\right] & \text{if } \hat{\lambda}^2_{ay}(l,k) > |\hat{D}^2(l-1,k)| \\ |\hat{D}^2(l-1,k)| & \text{otherwise.} \end{cases} \tag{2.15}$$

The $|\hat{D}^2(l,k)|$ represents the noise power estimate while α, β and γ are the constants selected experimentally.

Drawbacks

The major drawback in the spectral minima tracking method is that the non-linear tracking used to estimate the noise power that has a continuous smoothing without differentiating speech presence and absence periods. As a consequence, the estimator strongly follows the speech power spectrum and as a result the noise estimate increases whenever the noisy speech power spectrum increases irrespective of the changes in noise power level. This noise over-estimation due

to leakage of speech power into noise causes large amount of distortion in speech due to over-estimation as clearly shown in Fig. 2.3. This noise over-estimation is considerable for larger input SNR conditions, i.e., $> 5dB$, as noisy speech consists mostly the speech components. This can be noticed from Fig. 2.3 where the estimator provides over-estimation mainly occurs in the high input SNR segments (0-3s, & 6-9s).

2.3.3 Minimum Statistics (MS)

To avoid the noise over-estimation the minimum statistics (MS) method is proposed in [60], which involves the optimally smoothed noisy spectral power estimate and the analysis of the statistics of the spectral minima. It is based on the principle that the power level of the noisy speech often decays to the noise spectral power. Therefore, by tracking the minimum of the noisy speech spectrum, the noise estimate can be achieved. The key improvement here over spectral minima tracking method [59] is that it does not use a fixed smoothing factor α as used in Eq. (2.14) but a time and frequency dependent smoothing parameter, $\alpha(l,k)$. This time-frequency varying smoothing parameter is derived by minimizing the mean squared error (MSE) between the smoothed power spectrum $\hat{\lambda}^2_{ay}(l,k)$ and the noise estimate $|\hat{D}(l,k)|$, as

$$E\left[\left(\hat{\lambda}^2_{ay}(l,k) - |\hat{D}(l,k)|^2\right)|\hat{\lambda}^2_{ay}(l-1,k)\right], \tag{2.16}$$

where,

$$\hat{\lambda}^2_{ay}(l,k) = \alpha(l,k)\hat{\lambda}^2_{ay}(l-1,k) + (1-\alpha(l,k))|Y^2(l,k)|. \tag{2.17}$$

Substituting Eq. (2.17) into Eq. (2.16) and setting the first derivative to zero yields the optimum value for $\alpha(l,k)$ to

$$\hat{\alpha}_{opt}(l,k) = \frac{1}{1+\left(\frac{\hat{\lambda}^2_{ay}(l-1,k)}{|\hat{D}(l,k)|^2}|\right)^2}. \tag{2.18}$$

Although this $\hat{\alpha}_{opt}(l,k)$ provides satisfactory results and MS method offers significant improvement over Doblinger's spectral minima tracking method [59] in

terms of noise over-estimation, the estimated noise lags behind the true noise estimate and therefore, a correction factor α_c was suggested. Finally, the final time-varying smoothing parameter by using the suggested correction factor is given by

$$\hat{\alpha}_{opt}(l,k) = \frac{\alpha_m \alpha_c}{1 + \left(\frac{\hat{\lambda}^2_{ay}(l-1,k)}{|\hat{D}(l,k)|} \right)^2}, \tag{2.19}$$

where, α_m=0.96. The algorithm produces a noise estimate by distinguishing well

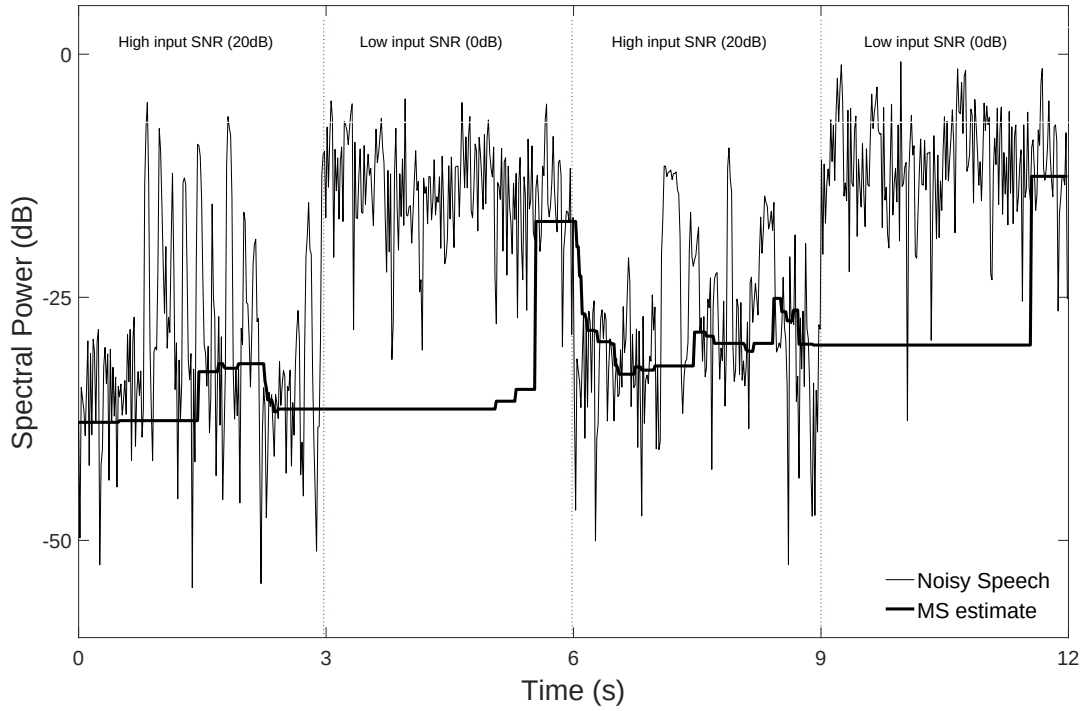


Figure 2.4: Plot of noisy speech power with MS estimate by using two input SNR conditions (0dB & 20dB) of babble noise at a given frequency 250Hz.

between an increase in noise power and an increase in speech power. However, the window length must be large enough to include the peaks of speech activity and short enough to follow the sudden noise variations.

The key improvement in this method is that, it does not use a fixed smoothing parameter as used in spectral minima tracking [59] method by providing an adaptive time-varying smoothing parameter as given in Eq. (2.19). The noise power estimated by MS method is shown in Fig. 2.4.

Drawbacks

The MS method mainly suffers by tracking capability as changes in noise estimate is delayed for any sudden noise spectral changes. This is because each window frame is divided into two sub-windows for increasing the estimation accuracy but, it requires larger memory size for processing to the window frame. Note that, the window length must be large enough to include the peaks of speech activity and short enough to follow sudden noise variations. Therefore, reducing the frame size in MS method, the delay problems arises because the time taken for processing each frame increases (almost double) largely.

Moreover, the estimator relies on the recursively updated noisy power given in Eq. (2.17), which means if the spectral minima for a given frame is unchanged, the estimator will fail to update the noise which causes the wrong estimation of the noise spectrum. This happens especially during low input SNR conditions (0dB time segments).

2.3.4 Minima Controlled Recursive Averaging (MCRA)

In a minima controlled recursive averaging (MCRA) method [62], the noise estimate is updated by averaging the past spectral values of noisy speech that is controlled by time-frequency dependent smoothing parameters. The calculation of these parameters is based on the speech presence probability (SPP) in each frequency bin, separately. The SPP is obtained by comparing the ratio of the noisy power spectrum to its local minimum against a fixed threshold. The binary hypotheses used in SPP in this method is as follows

$$\begin{aligned} H_0: &\quad Y(l,k) = |D(l,k)|, \\ H_1: &\quad Y(l,k) = |X(l,k)| + |D(l,k)|. \end{aligned} \tag{2.20}$$

The hypotheses H_0 represents the noise-only periods , whilst H_1 tells about speech present period similar to VAD. Based on these hypotheses, the noise estimate is sought, as

$$\begin{aligned} H_1: &\quad |\hat{D}(l,k)| = |\hat{D}(l-1,k)|, \\ H_0: &\quad |\hat{D}(l,k)| = \hat{\alpha}_d |\hat{D}(l-1,k)| + \big(1-\hat{\alpha}_d\big)|Y(l,k)^2|, \end{aligned} \tag{2.21}$$

where, $\hat{\alpha}_d$ is the SPP based smoothing parameter. In this method, the concept of introducing the binary hypotheses is that, the noise estimate is updated only when a noise-only period (H_0) is detected, that is similar to the methods based on the voice activity detection.

Speech Presence Probability (S_{pp})

The speech presence probability (S_{pp}) in each frequency bin is derived by using the VAD decision made from the ratio of the noisy speech power spectrum to its local minimum to a given threshold. The local minimum is found by smoothing the noisy speech power using Eq. (2.21) and finding the local minimum over a fixed window length, as

$$\begin{aligned} \lambda^2_{min}(l,k) &= \min\Big(\lambda^2_{tm}(l-1,k),\ \lambda^2_{ay}(l,k)\Big), \\ \lambda^2_{tm}(l,k) &= \lambda^2_{ay}(l,k), \end{aligned} \quad . \tag{2.22}$$

and a binary decision to find the speech present period (S_{pp}) is derived by

$$p'_p(l,k) = \frac{\lambda^2_{ay}(l,k)}{\lambda^2_{min}(l,k)}. \tag{2.23}$$

If S_{pp} is higher than a given threshold value, it is assumed that the speech is present in that particular frame, otherwise the frame contains only noise and by using Eq. (2.20), the noise estimate will be updated. The VAD decision in search of the speech present regions is given by

$$D_e(l,k) = \begin{cases} 0 \ \text{(noise-only)} & \text{if } \ p'_p(l,k) < \delta_{pp}, \\ 1 \ \text{(speech+noise)} & \text{elseif } \ p'_p(l,k) \geq \delta_{pp}. \end{cases} \tag{2.24}$$

Using the decision made by Eq. (2.24), the speech presence probability ($S_{pp}(l,k)$) is updated, as

$$S_{pp}(l,k) = \alpha S_{pp}(l,k) + (1-\alpha) D_e(l,k), \tag{2.25}$$

where, α is a smoothing parameter. Note that, the $D_e(l,k)$ provides the information about voice activity for a given frame and made a decision either speech active or speech inactive period (noise only). By using Eq. (2.25), the S_{pp} based

time varying smoothing constant yields

$$\hat{\alpha}_d(l,k) = \alpha_d + (1-\alpha_d)S_{pp}(l,k), \tag{2.26}$$

and the estimation of the noise can be achieved by using a time-varying smoothing parameter from Eq. (2.26), as follows

$$|\hat{D}(l,k)| = \hat{\alpha}_d(l,k)|\hat{D}(l-1,k)| + (1-\hat{\alpha}_d(l,k))|Y^2(l,k)|. \tag{2.27}$$

Since this method updates the local minimum recursively, the minimum value is updated by using the previous knowledge of the noisy speech power. This causes the miss-detection of the binary decision in finding the speech present/absent period. More specifically, for a poor SNR conditions ($\leq$7dB), the ratio in Eq. (2.23) fails to provide the accurate decision as it is compared with a fixed threshold. As a consequence, the speech presence probability based binary estimate $D_e(l,k)$ in Eq. (2.24), does not provide the correct decision. Besides that, to avoid falling to the global minimum, a temporary variable is estimated for each frame and equated to the noisy speech power spectrum at that frame. After that, the local minimum is calculated by using that temporary variable, and therefore, the tracking of the minima takes at most double frames to update the local minimum for increasing noise levels.

Improvements in MCRA (IMCRA)

The spectral adaptation of noise using the fixed threshold in [62], however lags behind, especially when the noise power increases abruptly. In [63], the ratio based SPP is replaced by the use of a conditional speech presence probability to recursively update the noise spectrum which yields an improvement in noise tracking. The derivation for noise estimate is based on the assumption that, the STFT of both speech and noise are Gaussian distributed. Then the conditional speech presence probability is given by

$$S_{pp}(l,k) = \left[1 + \frac{\Lambda}{1-\Lambda}\Big(1+\xi(l,k)\Big)e^{-\nu_k(l,k)}\right], \tag{2.28}$$

$$\nu_k(l,k) = \frac{\gamma(l,k)\xi(l,k)}{1+\xi(l,k)}, \tag{2.29}$$

where,

$$\Lambda = S_{pp}\Big[H_0|Y(l,k)\Big], \quad \xi(l,k) = \frac{\lambda^2_{ax}(l,k)}{\lambda^2_{ad}(l,k)}, \quad \text{and} \quad \gamma(l,k) = \frac{|Y^2(l,k)|}{\lambda^2_{ad}(l,k)}, \tag{2.30}$$

are the probability of noise-only, *a priori* SNR, and *a posteriori* SNR, respec-

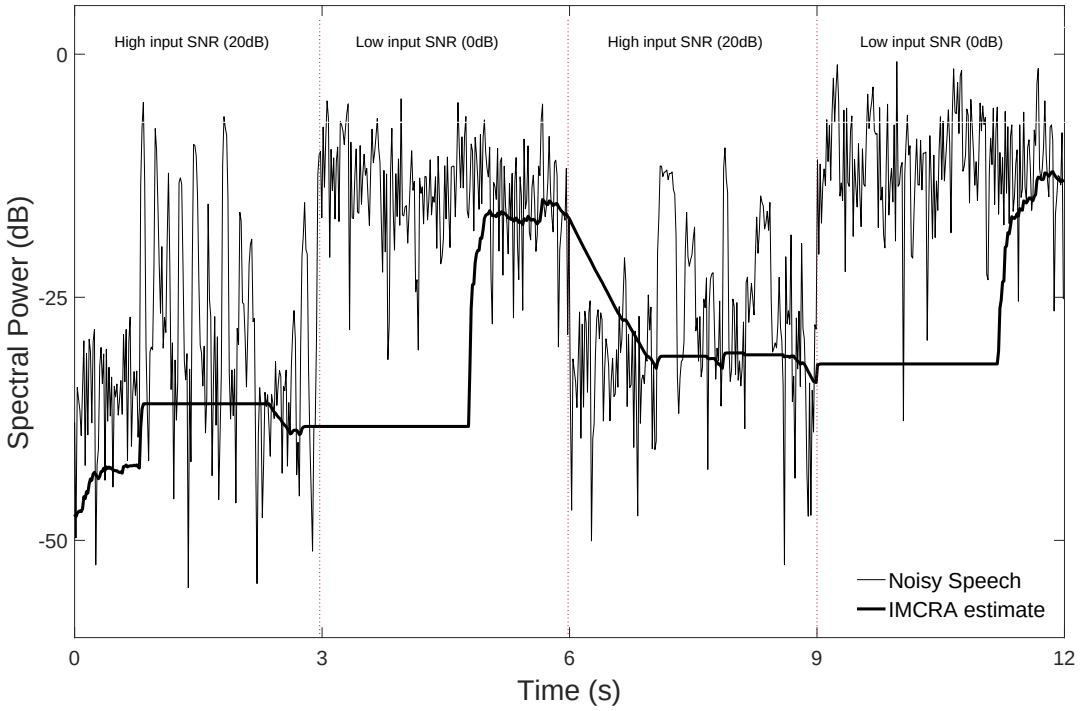


Figure 2.5: Plot of the true noisy speech power with IMCRA estimate by using two input SNR conditions (0dB & 20dB) of babble noise for a given frequency 250Hz.

tively. For the worst case scenario, the probability of speech presence and noise-only periods are half and therefore, Λ=1/2, is used. The noise estimate by using IMCRA method is shown in Fig. 2.5.

Drawbacks

Since, the estimation of noise in MCRA method is based on the SPP, it is quite difficult to identify the speech presence frame because both the speech and noise signals occupy almost same same spectral energy levels. This is especially the case during low input SNR conditions ($\leq$5dB). To update the noise power accurately, the SPP is replaced by the use of *a posteriori* probability in [63]. Although, this

method yields an improvement in noise tracking over [62], the computation of SPP is controlled by the minima values of a smoothed noisy power spectrum and therefore the noise estimate is influenced by tracking the minima of the spectrum. The problem arises, especially when the noise signal is non-stationary and has equal or higher energy than the speech signal. Additionally, the improved version of MCRA uses the principle of MS rule for tracking the minima of the spectrum, and therefore this IMCRA method faces similar problems of the estimation delay and processing time as compared to MS method [60]. Similar to MS method plotted in 2.4, the IMCRA updates the noise estimate with almost similar delay for low input SNR conditions as plotted in Fig. 2.5.

2.3.5 Statistical Model Based Noise Methods

Biased Noise Estimators

Recently, the minimum mean square error (MMSE) estimation based noise techniques have been developed [39, 65, 66]. In these methods, the mean square error (MSE) between the true and estimated noise spectrum is minimized by using the Bayesian based MMSE estimation.

In [39], the noise estimate is achieved by utilizing the principle of expectation-maximization (EM). In this, the instantaneous noise power is estimated based on information from the incoming signal and the current estimated distribution parameters. By assuming the zero-mean, complex Gaussian distribution for noise, the instantaneous noise power using MMSE estimation yields

$$|\hat{D}^2(l,k)| = E\Big[|\hat{D}^2(l,k)| \mid |Y(l,k)|\Big], \tag{2.31}$$

$$= \frac{\xi(l,k)}{1+\xi(l,k)}\left[\frac{\gamma(l,k)+\xi(l,k)\Big(\xi(l,k)+1\Big)}{\gamma(l,k)\xi(l,k)\Big(\xi(l,k)+1\Big)}\right], \tag{2.32}$$

where,

$$\hat{\xi}(l,k) = \alpha\frac{\hat{X}^2(l\text{-}1,k)}{\lambda_{ad}^2(l,k)} + (1-\alpha)\left[\max\Big(\gamma(l,k)-1,0\Big)\right], \tag{2.33}$$

is the *a priori* SNR estimated by using the decision-directed approach [20]. Since, the *a priori* SNR in Eq. (2.33) uses the previous knowledge of the speech power,

the noise estimate lags behind the true noise and, that results in leakage of speech signal of large amplitudes into the noise (over-estimation). To overcome this problem, a biasing compensation factor B_c, is introduced empirically, as

$$B_c = \frac{1}{\hat{\xi}(l,k)+1}\left[\frac{\Psi e^{-\Psi}}{1-e^{-\Psi}}\right]. \tag{2.34}$$

However, the estimator manage avoids the over-estimation problem, but unfortunately, provides the under-estimation of the noise power.

In [65], an analytically derived bias compensation factor is introduced to compensate for this biasing effect, as

$$B_c(\hat{\xi}) = \frac{1}{\left((1+\hat{\xi})\gamma_{ig}\left(\frac{1}{\hat{\xi}+1}, 2\right) + e^{-\frac{1}{\hat{\xi}+1}}\right)}, \tag{2.35}$$

where, $\gamma_{ig}(\mathsf{x}, r)$ is the incomplete gamma function of x. The noise estimate is then obtained by multiplying the $B_c(\hat{\xi})$ with the expectation of the noise power $E\Big[|\hat{D}^2(l,k)| \mid |Y(l,k)|\Big]$, yields

$$|\hat{D}^2(l,k)| = B\Big(\xi(\hat{l,k})\Big)E\Big[|\hat{D}^2(l,k)| \mid |Y(l,k)|\Big]. \tag{2.36}$$

Fig. 2.6 plots the effect of the bias compensation factor as the *a priori* SNR

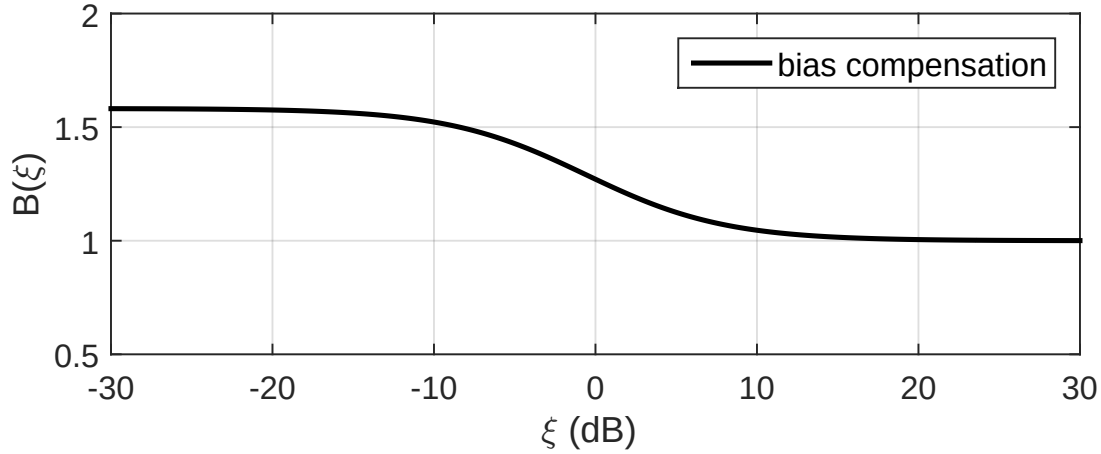


Figure 2.6: The effect of bias compensation factor (Eq. 2.35) on noise estimation.

estimate varies. It is clear that the biasing factor B reacts only when the noise is under-estimated as $B \geq 1$, especially for ξ greater than or equal to 10 dB,

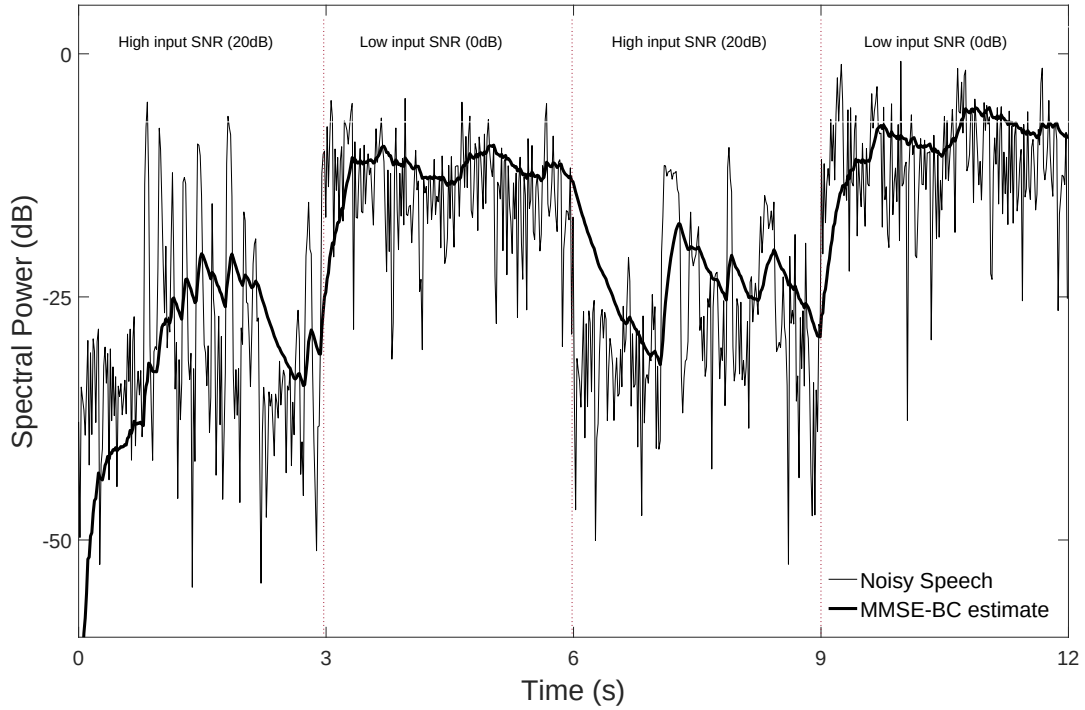


Figure 2.7: Plot of noisy speech power with MMSE-BC estimate by using two input SNR conditions (0dB & 20dB) of babble noise for a given frequency 250Hz.

whilst, the noise estimator is considered to be unbiased for over-estimation. The drawback with these methods [39, 65] is that both involve the estimation of the true a priori SNR.

Drawbacks

As we have seen, the MS and IMCRA methods suffers from the large estimation delay, the MMSE-BC method achieves good noise tracking comparatively. However, this method requires a biasing compensation factor. The major problem in this method is that, whenever the speech power is equal to noise power the estimator assumes that the frame has noise power only and the speech signal is misinterpreted by noise. This is because when noise power is nearly equal to speech power, the *a priori* SNR estimate approaches to zero which is the case for noise-only frames and as a result the estimated noise power increases even when speech is present in that frame. It can be clearly seen from Fig. 2.7 that the noise over-estimation occurs largely for low input SNR conditions, e.i., 0dB.

Unbiased Noise Estimators

A noise estimate, that does not require bias compensation factor is proposed in [66]. This method replaces the *a priori* SNR estimate $\hat{\xi}(l,k)$ with a soft-SPP by using two hypotheses as given in Eq. (2.20). Assuming that both the speech and noise complex coefficients are Gaussian distributed, the soft-SPP is given by

$$S_{pp}(l,k) = \left[1 + \frac{\Lambda}{1-\Lambda}\left(1+\xi_{H_1}\right)e^{-\left(\frac{\xi_{H_1}}{1+\xi_{H_1}}\gamma(l,k)\right)}\right], \tag{2.37}$$

where, the ξ_{H_1} is a fixed *a priori* SNR 15dB selected experimentally. Although, this soft-SPP based method provides better noise estimate compared to MS [60] and IMCRA [63] with faster noise estimation, results the similar performance as achieved by [65]. Additionally, for a sudden change in non-stationary noise power, i.e., in both high SNR to low SNR conditions (3s to 6s and 9s to 12s) it fails to adopt quick changes in noise statistics as shown in Fig. 2.8.

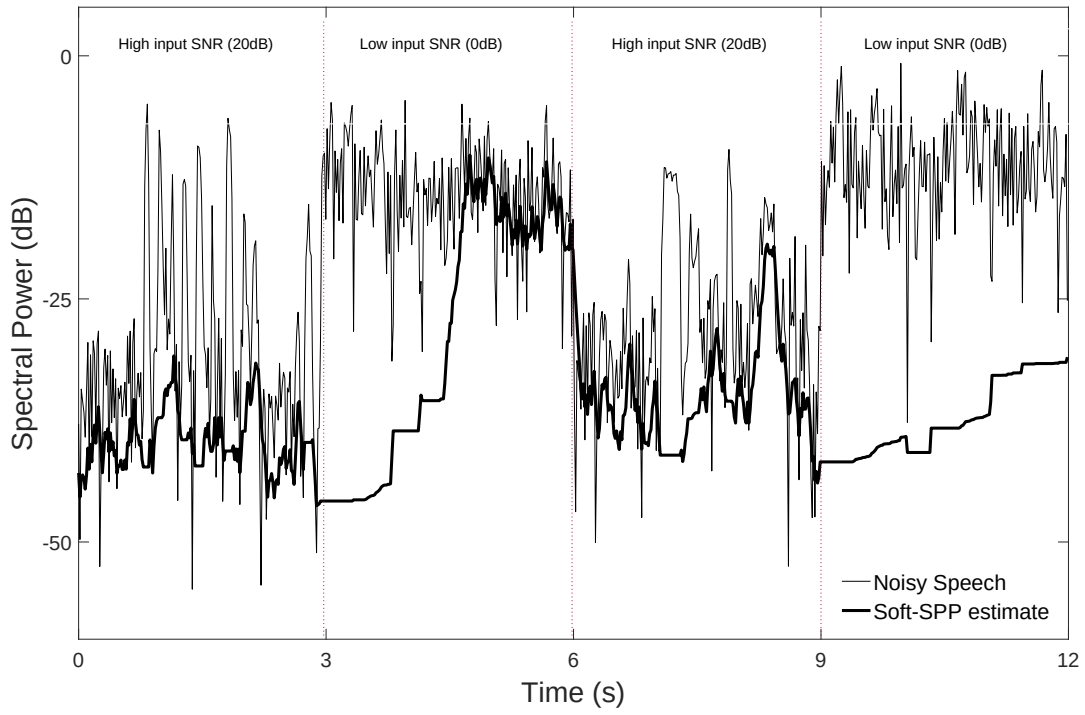


Figure 2.8: Plot of noisy speech power with soft-SPP estimate by using two input SNR conditions (0dB & 20dB) of babble noise at a given frequency 250Hz.

The method [66] proposes a computationally efficient algorithm for estimating the noise spectral power. As the soft-SPP method [66] uses the Gaussian assumption for noise distribution, [67] uses the sigmoid function to find the opti-

mal solution for the soft-SPP $S_{pp}(l,k)$, as follows

$$S_{sig}(l,k) = \frac{1}{1+exp\Big(-A\Big[\gamma(l,k)-B\Big]\Big)}, \tag{2.38}$$

where,

$$\begin{aligned} 0.94 \leq A = \tfrac{\xi_{H_1}}{1+\xi_{H_1}} \leq 0.98, \qquad \text{and} \\ 3.00 \leq B = \Big[\tfrac{1+\xi_{H_1}}{\xi_{H_1}}\Big]\log\big(\tfrac{\Lambda}{1-\Lambda}\big) \leq 4.23, \end{aligned} \tag{2.39}$$

are the slope and mean of the sigmoid function, respectively. Moreover, the *a posteriori* SPP in this method is categorized into three deferent probability levels by

$$S_{pp} = \begin{cases} \mathcal{P}_1 & : \text{less likely speech presence} \quad 0.30 \geq S_{sig}, \\ \mathcal{P}_2 & : \text{more likely speech presence} \quad 0.30 < S_{sig} \leq 0.60, \\ \min\big[\mathcal{P}_3, S_{sig}\big] & : \text{most likely speech presence} \quad 0.60 \leq S_{sig}, \end{cases} \tag{2.40}$$

and the exponential time-frequency smoothing constants $\mathcal{P}(j{=}1,2,3)$ are given by

$$\mathcal{P}_1 = e^{\frac{-2.2S}{t_1 F_s}}, \quad \mathcal{P}_2 = e^{\frac{-2.2S}{t_2 F_s}}, \quad \textit{and} \quad \mathcal{P}_3 = e^{\frac{-2.2S}{t_3 F_s}}, \tag{2.41}$$

where, $t1 < t2 \leq t3$ denotes the averaging time constant with frame rate S and sampling frequency F_s. As [66], suggests a fixed value for the *a priori* SNR ξ_{H_1}=15dB, the sigmoid function based noise method argued by saying that any value from 12dB to 18dB can be used for ξ_{H_1} as the noise estimator achieves similar results. Fig. 2.9, clearly suggest that the sigmoid based noise estimator [67], has better noise tracking and improvement in terms of speech quality and intelligibility is perceived with the flexibility and less computational complexity of the noise estimator compared to [66].

However, the key assumption in the noise estimation techniques is that the noise spectral coefficients are assumed to follow the Gaussian distribution [20, 38, 39, 63, 65, 66]. Whilst the Gaussian assumption may be sufficient for stationary noise, it may not be the case for non-stationary noise, such as babble or street noise having a time-varying noise probability distributions. Since, the single-microphone based signal consists a mixture of speech and noise signals, separat-

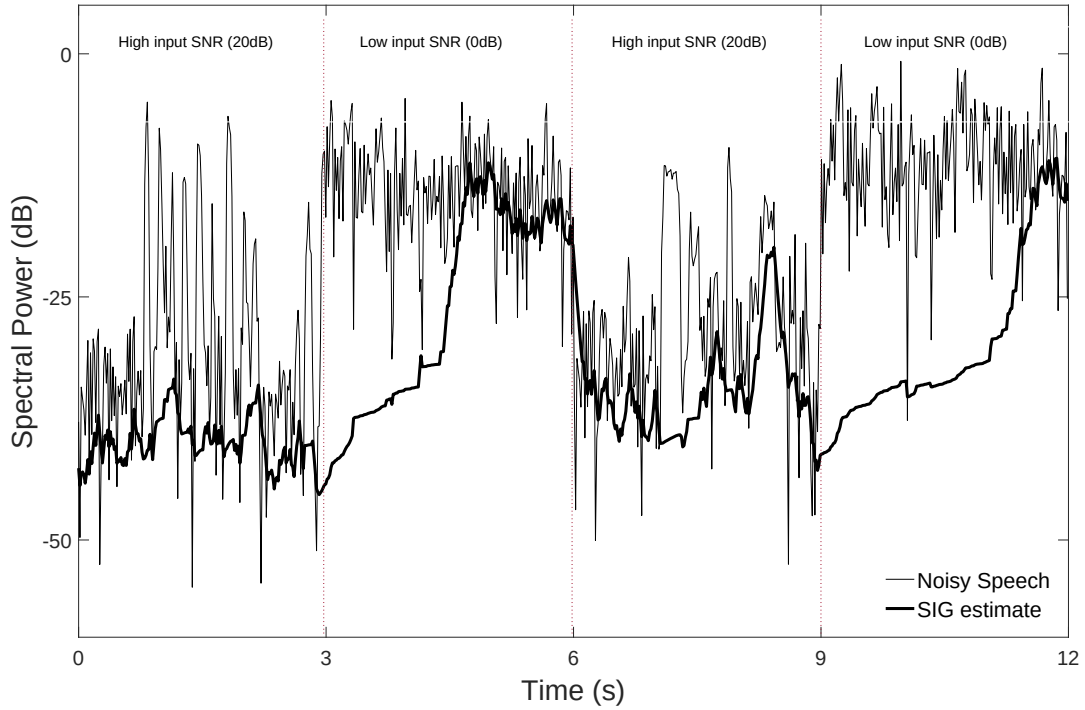


Figure 2.9: Plot of noisy speech power with SIG estimate by using two input SNR conditions (0dB & 20dB) of babble noise at a given frequency 250Hz.

ing both signals are a difficult task. Also both signals have different probability distributions in spectral domain and, the noise such as in a restaurant, or in a social gathering has a time-varying nature, thus characterization of noise probability distribution by using the Gaussian assumption is not adequate and, results remain inconclusive as no single density function can represent the different real world noise.

The above complication can be reduced by using Bayesian estimation theory that minimizes the Baye's risk function, which includes a posterior probability model of the unknown parameters (given from the observation vector) and a cost error function [20]. The posterior probability density function (pdf) depends on how relatively the noise pdf is peaked, i.e., the likelihood pdf depends on the posterior pdf. The more peaked the noise pdf, the larger the estimation error will be, and as a result, the greater the influence on the outcome of the noise estimation process. Conversely, a uniform pdf will have no influence on the estimation [43].

This observation leads to use the modulation DFT coefficients. Since energy from two different signals (speech and noise) is largely non-overlapping in the

modulation domain [28]. This suggests that estimating individual information (separating two signals) in the modulation domain is relatively easier [36], and therefore, the modulation spectrum may assist in the demarcation of speech and noise recorded from a single-microphone.

Therefore, for the modulation based speech enhancement, the following section (2.4) describes the complete structure of the modulation transform focusing on the achievements of various single-channel speech enhancement techniques by using the modulation domain processing.

2.4 Modulation domain Speech Enhancement

The modulation domain has been reported to be a better alternative to the frequency domain, in particular for a single-microphone speech enhancement, as the "modulation-frequency" is highly correlated to the speech intelligibility. There are many substantial evidences supporting that, the modulation-frequency concept is useful for describing, representing and modifying audio signals as low-frequency modulators can represent the audio signals.

Zadeh [84] was the first who represented a signal by using a 2-dimensional frequency model, where the second dimension for frequency analysis was achieved by the transform of the time-variation of the acoustic frequency. Later, several studies have shown the importance of low-frequency modulators for speech reception [85]. For example, Viemeister in [86–88] represents the auditory system by using an empirical function called Temporal Modulation Transfer Function (TMTF) in which, the TMTF does the modulation thresholding that helps to detect the amplitude-modulation (AM) of a sinusoidal signal as a function of the modulation frequency. For an effective model of modulation masking and detection, a modulation filter-bank is applied to the signal [31, 32]. In their models, they assumed that for the modulation frequency 0-10Hz modulation filters are uniformly spaced with a bandwidth of 5Hz. In recent years, this concept of modulation filter-bank in the auditory system received considerable attentions in the field of speech enhancement [29, 30, 33, 35, 89–95].

More importantly, the correlation of modulation frequencies with linguistic in-

formation of speech (intelligibility) has been investigated in [35,92]. They applied both low-pass and high-pass filters to the temporal envelopes of the acoustic frequency sub-bands and, found that the frequency between 4Hz to 16Hz are highly correlated to speech intelligibility. Subsequently, [30] conducted the similar experiment and argued by saying that applying a band pass filter between 1Hz to 16Hz to modulation frequencies does not impair speech intelligibility. The argument provided that the acoustic spectrum only provides the knowledge about vocal tract shape whilst, the modulation envelope shows the changes in vocal tract with respect to time. These vocal track changes convey most of the linguistic information of the speech (intelligibility). Besides that, the lower limit of 1 Hz could be from the fact that the slow vocal tract changes do not provide much linguistic information while the upper limit of 16 Hz is due to the physiological limitation on how fast the vocal tract is able to change with time [27]. There are many modulation based speech applications such as in speech coding [23, 24, 96], speech and speaker recognitions [97–100], as well as in speech enhancement [27,34,101–106] have found a growing interest in the field of speech processing.

In a more systematic way, [95] represented the acoustic frequency as the axis of the first STFT (acoustic transform) of the input signal and the modulation frequency as an independent variable of the modulation transform (second STFT). Simply, the modulation spectrum is the STFT of the time series of the acoustic spectrum for a given acoustic frequency. The related work of enhancing the single-channel based noisy speech by using the modulation domain framework is followed by [27,34,102,103].

In the following section 2.4.1, we provide the details of a simplified modulation transform for single-channel speech enhancement as shown in Fig. 2.10.

2.4.1 Modulation Transform

The modulation noisy speech spectrum can be achieved by applying the secondary AMS framework (modulation transform) of the time variation of the acoustic frequency as explained in [95], where the primary AMS framework (acoustic-transform) gives the frequency domain spectrum of the time-domain noisy speech

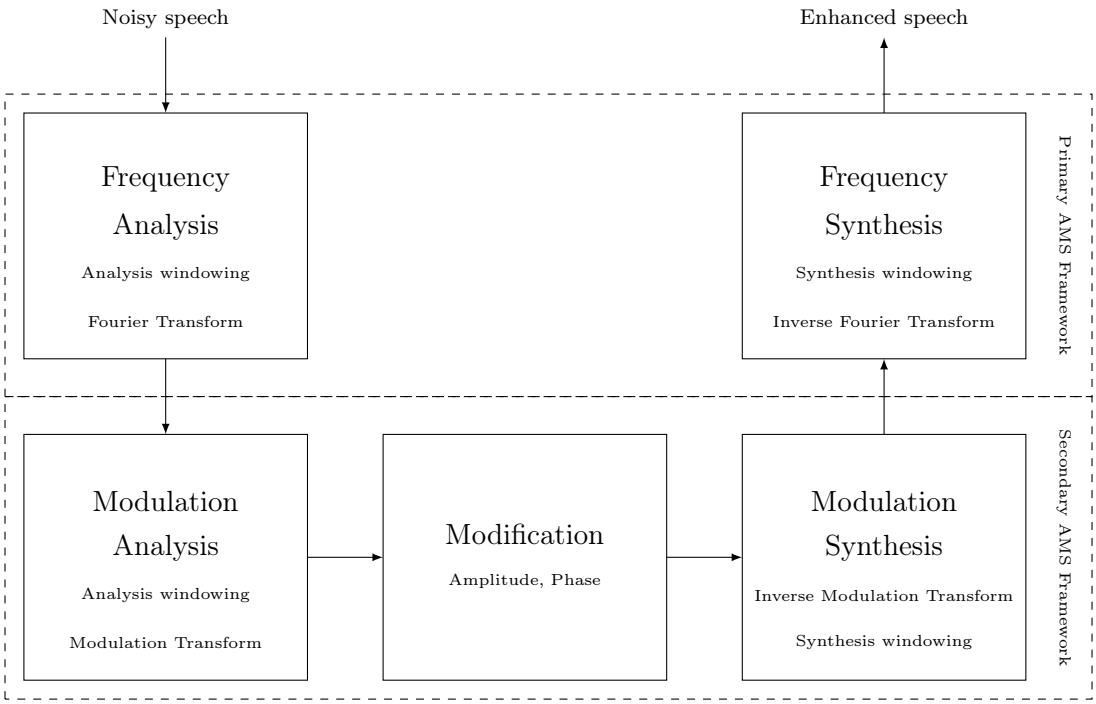


Figure 2.10: Generalized block diagram of a modulation transform based single-channel speech enhancement.

signal, whilst the modulation spectrum achieved by applying the modulation transform to the time series of an acoustic spectrum for a given frequency band. Therefore, the short-time modulation spectrum is a function of time, acoustic frequency, and the modulation frequency.

As we have already achieved the acoustic magnitude spectrum in Eq. (2.4), applying the modulation transform to each acoustic frequency index k of the noisy speech magnitude spectrum, gives

$$Z(\tau, k, m) = \sum_{l=0}^{M-1} |Y(l + \tau P, k)| w_{ms}(l)\ e^{-j\frac{2\pi ml}{M}}, \tag{2.42}$$

where, $|Y(l, k)|$ is the acoustic domain noisy speech magnitude spectrum, m represents the modulation frequency index (0, 1, . . ., M-1), τ is the modulation time index, M is the modulation FFT (MFFT) size, P is the modulation frame shift (MFS), and $w_{ms}(l)$ is the modulation analysis window function. Note that, both MFFT and MFS are given in terms of the acoustic sampling frequency. For example, if the primary AMS framework uses 32ms to frame the noisy speech sampled at 8kHz, a 512-point AFFT size [1] is achieved, where each acoustic fre-

[1]The frames in both frequency and modulation domains were padded with zeros to double

quency bin is sampled at 62.5 Hz and has a bandwidth of 15.625Hz.

Since, it is assumed that the noise is uncorrelated and additive in the acoustic domain, Eq. (2.4) shows the linearity properties of acoustic DFT coefficients, as

$$Y(l,k) = X(l,k) + D(l,k), \tag{2.43}$$

and, and noisy magnitude spectrum

$$|Y(l,k)| = |X(l,k)| + |D(l,k)|. \tag{2.44}$$

Similarly, applying the linearity property of noisy speech DFT coefficients to modulation domain, yields

$$Z(\tau,k,m) = \mathcal{X}(\tau,k,m) + N(\tau,k,m), \tag{2.45}$$

where, the complex-valued DFT spectral coefficients of noisy, clean and noise signals $Z(\tau,k,m)$, $\mathcal{X}(\tau,k,m)$ and $N(\tau,k,m)$ are

$$\begin{aligned} Z(\tau,k,m) =& |Z(\tau,k,m)|\ e^{j\angle Z(\tau,k,m)}, \\ \mathcal{X}(\tau,k,m) =& |\mathcal{X}(\tau,k,m)|\ e^{j\angle \mathcal{X}(\tau,k,m)}, \quad \text{and} \\ N(\tau,k,m) =& |N(\tau,k,m)|\ e^{j\angle N(\tau,k,m)}. \end{aligned} \tag{2.46}$$

Further the modulation magnitude spectrum in the modulation domain can be written as

$$|Z(\tau,k,m)| = |\mathcal{X}(\tau,k,m)| + |N(\tau,k,m)|, \tag{2.47}$$

where the relation of their respective modulation domain variances holds

$$\lambda^2_{my}(\tau,k,m) = \lambda^2_{mx}(\tau,k,m) + \lambda^2_{mn}(\tau,k,m), \tag{2.48}$$

the length throughout this thesis work, that results in the frequency and modulation FFT sizes composed of 2K and 2M, respectively.

where,

$$\begin{aligned}\lambda^2_{my}(\tau,k,m) &= E\big[Z^2(\tau,k,m)\big],\\ \lambda^2_{mx}(\tau,k,m) &= E\big[X^2(\tau,k,m)\big], \quad \text{and}\\ \lambda^2_{mn}(\tau,k,m) &= E\big[N^2(\tau,k,m)\big],\end{aligned} \tag{2.49}$$

are the modulation domain variances of noisy speech $Z(\tau,k,m)$, clean speech $\mathcal{X}(\tau,k,m)$ and noise $N(\tau,k,m)$, respectively.

After the modification stage of either magnitude or phase spectrum, the modified speech spectrum in the modulation domain $\hat{\mathcal{X}}(\tau,k,m)$, as shown in Fig. 2.11 is used to estimate the acoustic speech magnitude spectrum $|\hat{X}(l,k)|$, by applying the inverse modulation transform followed by least-square overlap-add method with modulation synthesis window [70], as

$$\hat{X}(l,k) = \sum_{\tau}\left\{ w_{ms}(l-\tau P)\sum_{m=0}^{M-1}\hat{\mathcal{X}}(\tau,k,m)\; e^{j\frac{2\pi(l-\tau P)m}{M}}\right\}. \tag{2.50}$$

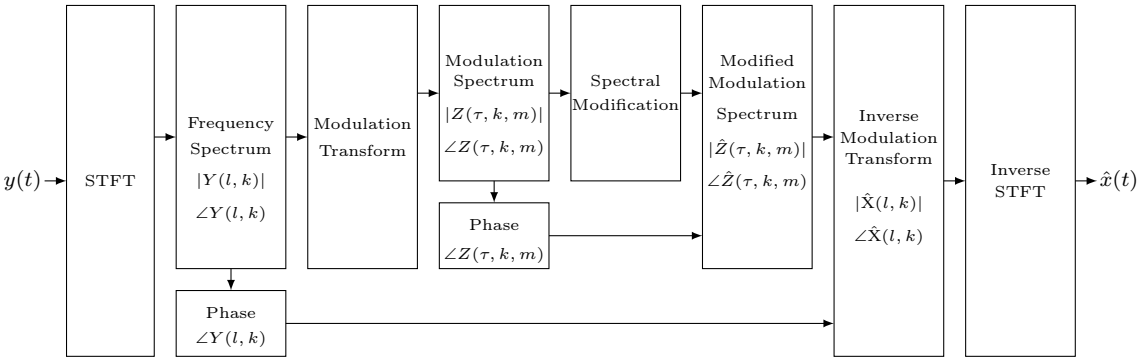


Figure 2.11: The modulation-transform representation for the single-channel speech enhancement.

The modified acoustic speech spectrum $\hat{\mathbf{X}}(l,k)$ can then be found by combining the modified magnitude $|\hat{X}(l,k)|$ of speech and phase spectra $\angle Y(l,k)$. Finally, the enhanced speech is synthesized by applying inverse STFT to modified acoustic spectrum of speech $\hat{X}(l,k)$, followed by the overlap-add synthesis given as

$$\hat{x}(t) = \sum_{l}\left\{ w_{as}(t-lS)\sum_{k=0}^{K-1}\hat{X}(l,k)\; e^{j\frac{2\pi(t-lS)k}{K}}\right\}, \tag{2.51}$$

where, $\hat{x}(t)$ is the recovered speech signal in time domain and $w_{as}(t)$ is the hamming window function.

2.5 Summary

This chapter reviewed the developments that have been made in the field of single-channel speech enhancement based noise estimation. Based on the literature, it is found that the VAD based noise methods fail to differentiate the noise and speech spectra correctly, specially at poor SNR conditions ($\leq 5dB$). This results in either noise under-estimation (residual noise) or over-estimation (speech distortion) and the noise estimator suffers by the loss of speech intelligibility.

On the other hand, non-VAD based noise methods achieve better noise estimates as compared to the VAD based noise methods but in a high noise conditions ($\leq 5dB$), they fail to track the highly non-stationary noise spectrum. The reason may be that these methods use the Gaussian assumption for noise spectral coefficients. Whilst the Gaussian assumption may be sufficient for stationary noise, it may not be the case for highly non-stationary babble or heavy street noise signals. Moreover, each noise method has it's own limitations, e.g., the Minimum Statistics (MS), MCRA and IMCRA methods suffer by large estimation delay especially when the noise signal is highly non-stationary and has equal or higher energy than the speech signal. Whilst, MMSE-BC method needs and additional biasing compensation factor to avoid the speech power leakage in to the noise power.

In the next chapter, problem of noise distribution function will be considered by focusing on all time-varying noise signals. Moreover, the properties of modulation domain will be explored so that the problem associated with the aforementioned noise estimation methods can be tackled by the proposed noise estimator by using best suited noise distribution function in the modulation domain.

$\sim * \sim$

Chapter 3

Modulation Domain Noise Modeling

If you want to shine like a sun, first burn like a sun.

–Dr. APJ Abdul Kalam.

3.1 Introduction

As mentioned in chapter 2, the Gaussian assumption for noise spectral coefficients is not adequate for non-stationary noise such as restaurant noise or babble noise where the noise spectral coefficients continuously change with time.

Studies in [107, 108] suggest that most semi-stationary environmental noise signals such as car noise are approximately super-Gaussian distributed and can be better fitted with a Laplacian density. Similarly, non-stationary noise signals are assumed to be better fitted with the Gamma density function [60]. Later in [109], the noise DFT amplitudes are modeled by using Rayleigh and Laplacian densities for stationary white noise, fan noise and babble noise. The fitted histograms of these noise types show that the deviation of the measured noise histogram from the Rayleigh density is lower compared to the Laplacian density. However, results remain inconclusive as no single density function can characterize the different real world noise signals.

The above complication can be reduced by using the Bayesian estimation the-

ory [20,21]. This is because the optimal estimators can be obtained by minimizing the Baye's risk function, which includes a posterior probability model of the unknown parameters (given from the observation vector) and a cost error function. The posterior probability density function (pdf) depends on how relatively the noise pdf is peaked, i.e., the likelihood pdf depends on the posterior pdf. Generally, the more peaked the noise pdf, the larger the estimation error will be, and as a result, the greater the influence on the outcome of the noise estimation process. Conversely, a uniform pdf will have no influence on the estimation [43]. Clearly, the closer the model to the actual distribution is the crux of the problem.

This chapter sets out to investigate the use of modulation domain for noise modeling and estimation. Schimmel in his dissertation [28] reports that the energy from two different signals (e.g., speech and noise signals) in the modulation domain is largely non-overlapping. This is also supported by psychoacoustic research, which indicates that the human auditorial system segregates sound in the modulation domain [29]. This suggests greater demarcation between the noise and the speech in the modulation domain. Also, various research, which dates back to the early 90s, indicate that modulation domain processing often results in higher intelligibility of speech [27, 30–34]. Moreover, the intelligible components of the speech signal are mostly confined to the modulation frequency band of 1Hz to 16Hz and therefore processing can be made to concentrate on the relevant bands [35, 36].

This chapter derives a Bayesian based noise estimation method in the modulation domain for speech intelligibility improvement. The first part of the chapter entails a study on the suitability of the modulation domain in Bayesian estimation. The study shows that the modulation domain provides a better matching between the various real world noise densities with well established density models compared to the conventional frequency domain. Importantly, the investigation also reveals that the spectral variation in the modulation domain is more uniform compared to the acoustic domain irrespective of the type of noise and SNR levels. As mentioned, a less peaked pdf will result in a smaller estimation error and as such, modulation domain is highly suitable. The study found that the modulation based Gamma density consistently provides the best pdf model for various

stationary and non-stationary noise. The Gamma density function is then used to derive the modulation based noise estimator by using a minimum mean square error (MMSE) based Bayesian estimator. We also show that the proposed modulation based noise estimator the noise is bias free. This is a direct consequence from the more accurate representation of the modulation based Gamma density function. Comprehensive experimental results show that the modulation based proposed noise method contributes to improving the speech intelligibility.

3.2 Modulation Domain Characteristics

3.2.1 Spectral Characteristics

As mentioned previously, various physiological and psychoacoustic findings show modulation domain processing highly correlates with improvement in speech intelligibility. However, the spectral stationarity changes by changing the window length and therefore in search of a suitable noise density function in the short-time modulation domain, the spectral characteristics (noise statistics) in terms of stationarity has to be explored. For this, the spectral characteristics of the modulation DFT coefficients against the acoustic DFT coefficients have been compared by using the excess kurtosis measure. The kurtosis measures the degree to which a distribution is more or less peaked than a normal distribution. Positive kurtosis indicates a relatively peaked distribution whilst the negative kurtosis indicates that the distribution is relatively flat [110]. In this investigation, the speech signals degraded by three different noise types at five input SNR levels are considered and the acoustic DFT coefficients are achieved by using a 512-point AFFT (50% AFS) whilst modulation envelope by using 512-point AFFT (6.25% AFS) for primary AMS framework and a 32-point MFFT (50% MFS) for modulation transform.

Results from Table 3.1 show that the modulation domain provides more uniformity (less impulsive) of the noisy speech compared to the acoustic spectral variation irrespective of the type of noise and SNR levels. Also, the more predictable trend in the kurtosis measure is observed as a function of SNR compared to the measure in the acoustic domain. Thus, the probability of miss-detection

Table 3.1: Comparison of mean kurtosis scores between the Acoustic and modulation domains based noisy speech DFT coefficients by using a wide range of input SNRs. The mean kurtosis score of clean speech is given for reference.

Input	white noise		factory noise		Babble noise		Street noise	
SNR	Acoustic	Modulation	Acoustic	Modulation	Acoustic	Modulation	Acoustic	Modulation
0	14.4457	3.8677	14.8953	2.3137	13.4030	2.5595	14.1535	2.4039
5	14.4023	2.9000	14.1403	2.2279	14.3958	2.3159	14.4871	2.2827
10	14.3740	2.4119	14.3053	2.2047	14.4175	2.2172	14.4615	2.2303
15	14.3629	2.2718	14.3352	2.1930	14.3873	2.1829	14.4203	2.2032
20	14.3587	2.2090	14.3452	2.1888	14.3710	2.1746	14.3932	2.1916

clean	Acoustic	Modulation
speech	14.35	2.10

(wrong-estimate) due to the heavy tails in spectral changes of non-stationary noise will be reduced in the modulation domain. Interestingly, studies on modulation domain do not have sufficient knowledge of spectral variations and therefore this spectral stationarity may be one of the main reasons why the modulation domain as a better alternative to the acoustic domain for speech enhancement, especially, in terms of speech intelligibility improvements as shown in Table 3.4 to 3.7. Additionally, this may assist in the demarcation of the speech and noise signals more effectively [27, 30, 35, 111].

3.2.2 Noise Distribution Model

A vast majority of speech enhancement methods assume the noise signal to be Gaussian distributed [39, 63, 65, 66, 109]. However, in the real world, noise can manifest itself in many ways, stationary or non-stationary, and as a result, it will differ from the aforementioned Gaussian assumption [107, 112]. Here, we further investigate the noise distribution modeling in both acoustic and modulation domains. The reason for this investigation is to ascertain if modulation domain provides better matching in terms of various real world noise densities with well established density models. A comparison is then made with the actual histogram of the noise DFT amplitudes with the Gaussian, Rayleigh, and Gamma models, where the histogram plotted the noise DFT amplitudes and extended by

distribution fitting with these distribution models.

For a given random variable n, the Gaussian, Rayleigh, and generalized Gamma probability density functions (pdf) can be represented by

$$f_{Gauss}(n) = \frac{1}{\sqrt{2\pi\lambda_g^2}} e^{-\frac{n^2}{\lambda_g^2}}, \tag{3.1}$$

$$f_{Ray}(n) = \frac{n}{\sqrt{2\pi\lambda_r^2}} e^{-\frac{n^2}{\lambda_r^2}}, \tag{3.2}$$

$$f_{Gamma}(n) = \frac{1}{\Gamma(\nu)\beta^\nu} n^{\nu-1} e^{-\frac{n}{\beta}} \quad \forall \quad n > 0, \tag{3.3}$$

where λ_g, λ_r are the noise variances of the Gaussian and Rayleigh densities respectively. The constants β and ν are the scale and shape parameters of the Generalized Gamma density, respectively, and $\Gamma(\nu)$ is the Gamma function evaluated at $\nu \geq 0$. Experiments are conducted with stationary white noise along with semi-stationary factory noise and highly non-stationary babble and street noise types. Each noise signal has a length of 600s sampled at 8kHz. For the acoustic noise DFT amplitudes, AMS framework has been applied to noise signal by using 512-point FFT size (frame length-32 ms[1]) with a 50% frame shift.

As for modulation noise DFT amplitudes, primary AMS framework uses a 512-point FFT with a 6.25% frame shift and the modulation based noise DFT amplitudes are achieved by using a 32-point FFT with a 50% frame shift in modulation transform (secondary AMS framework). The achieved DFT amplitudes are normalized to a unit variance. For illustration purpose, the plotted histograms are compared with the Gaussian (3.1), Rayleigh (3.2), and Gamma (3.3) models by using the frequency bin 1kHz[1] and a modulation frequency of 12Hz. The goodness-of-fit test is also conducted by using the one-dimensional Kolmogorov-Smirnov (K-S) test [113].

The respective deviation measures between the histogram and given noise distribution models are shown in Fig. 3.1 to 3.4 in terms of their goodness-of-fit (closeness). From both the acoustic and modulation histograms (Figs. 3.1-3.4), it is clear that the Gamma density delivers the best approximations to the actual noise in both DFT domains. Interestingly, in Fig. 3.1.a, the Rayleigh

[1]A similar analysis is done by using all acoustic frequency bins ranging from 50Hz to 3.5kHz, while all modulation frequency bins for histograms fitting.

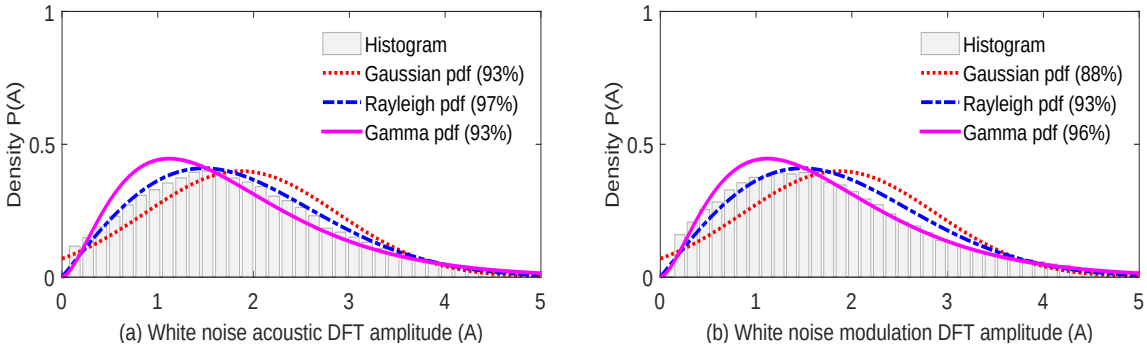


Figure 3.1: Histogram of white noise DFT amplitudes in (a) acoustic domain and in (b) modulation domain.

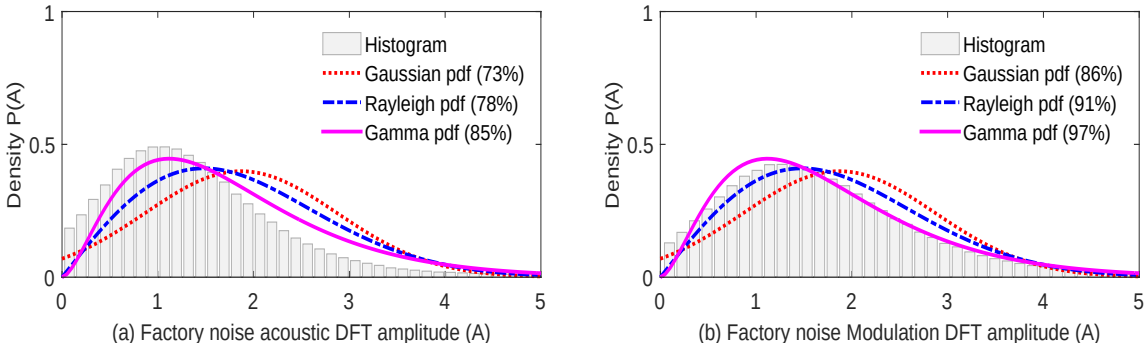


Figure 3.2: Histogram of factory noise DFT amplitudes in (a) acoustic domain and in (b) modulation domain.

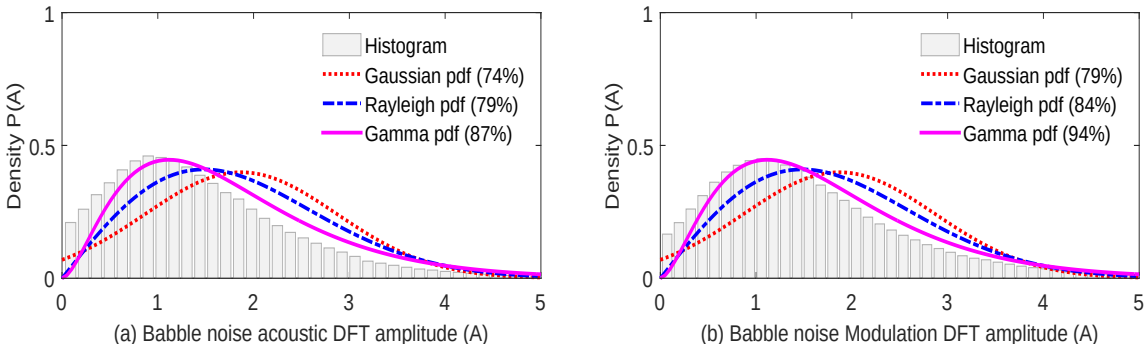


Figure 3.3: Histogram of babble noise DFT amplitudes in (a) acoustic domain and in (b) modulation domain.

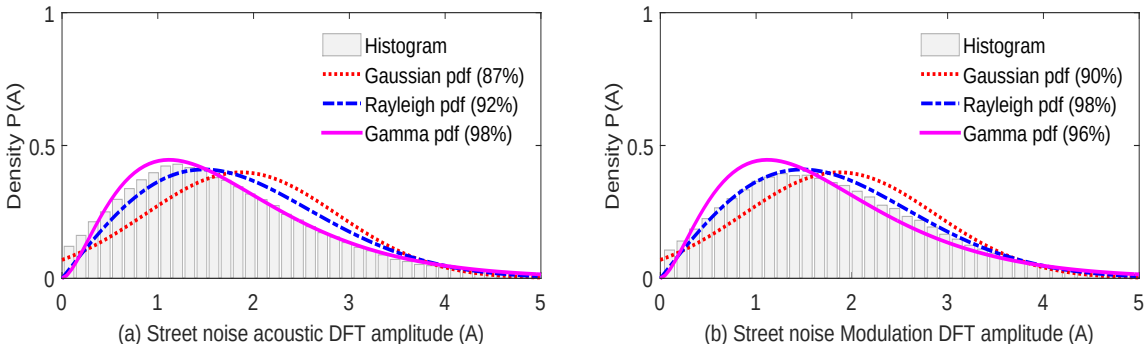


Figure 3.4: Histogram of street noise DFT amplitudes in (a) acoustic domain, and in (b) modulation domain.

density for white noise provides an almost perfect fitting (97%) to the actual noise histogram. This is because, the DFT coefficients (real and imaginary parts) of white noise tend to be Gaussian distributed in the acoustic domain [20, 109]. Contrary to white noise, if the speech is corrupted by practical noise such as the factory, cafeteria or street noise signals; the spectral amplitudes tend to be heavy tail due to rapid spectral fluctuations. For example, babble noise degrades the speech intelligibility more than the white noise does, because the vowel portion is mainly affected due to multiple speech components coming from neighboring speakers. As a result, the noise becomes speech-like when the input SNR is less than 7 dB [46]. This problem of heavy tail in real world noise signals especially in babble noise will be reduced in modulation domain which results the lower deviation by providing a closer fit which is clearly observed when modeling in the modulation domain compared to the acoustic domain.

Based on these findings, this chapter adopts the Gamma density for tracking the noise in both the acoustic and modulation domains as apposed to the conventional Gaussian assumption for all time varying noise signals.

3.3 Proposed Noise Estimation Method

Since the Gamma density yields the least deviation from the true noise distribution as given in Figs. 3.1 to 3.4, this section derives the Gamma density based noise estimation method by using the Bayesian approach in a minimum mean-square-error (MMSE) sense. In the sequel, we will consider the processing of a single modulation frame and, therefore, frame index is omitted.

Consider the Bayesian estimation of modulation based noise spectra, the posterior probability density function (pdf) $f_{N|Z}(n|z)$ of the noise signal N for a given noisy signal Z, can be written as

$$f_{N|Z}(n|z) = \frac{f_{Z|N}(z|n)\ f_N(n)}{f_Z(z)}, \tag{3.4}$$

where $f_{Z|N}(z|n)$ is the likelihood function generated by the discrete noise vector n, $f_Z(z)$ and $f_N(n)$ are the prior probabilities of the noisy and noise signals, respectively. Note that, $f_Z(z)$ is constant for a given noisy signal Z and has no

effect on the risk-minimization process. As mentioned earlier, the posterior pdf $f_{N|Z}(n|z)$ depends on the shape of these two functions, e.g., the peakedness of $f_{Z|N}(z|n)$ relative to $f_N(n)$. Strictly, the more peaked the posterior pdf is, the larger the estimation error will be, and this will ultimately influence the noise estimate. Conversely, a uniform pdf will have no influence [43].

The Bayesian risk function can be defined in terms of the average cost-error function $C(n,\hat{n})$, and $f_{N|Z}(n|z)$, as

$$\begin{aligned}\mathcal{R}(\hat{n}) &= E[C(n,\hat{n})]. \\ &= \int_n \int_z C(n,\hat{n}) f_{N|Z}(n|z) f_Z(z) dz dn.\end{aligned} \tag{3.5}$$

Note that $C(n,\hat{n})$ allows tuning in the form of spectral weighting so that the estimator can achieve the desired outcomes. By using the squared-error cost function $[C(n,\hat{n})=(n-\hat{n})^2]$, the Bayesian estimate is obtained by assuming that the risk function is differentiable and has a well-defined minimum. Differentiating the risk function (3.5) and setting the gradient to zero yields

$$\frac{\partial \mathcal{R}(\hat{n},z)}{\partial n} = -2\int_n \int_z (n-\hat{n}) f_{N|Z}(n|z) dz dn = 0. \tag{3.6}$$

Solving for the conditional expectation $\mathbb{E}[\hat{N}|Z]$ [20], we get

$$E[|\hat{N}|^2 \mid Z] = \frac{\int_0^\infty n^2 \; f_{Z|N}(z|n) f_N(n) dz dn}{\int_0^\infty f_{Z|N}(z|n) f_N(n) dz dn}, \tag{3.7}$$

where, $E[.]$ is the expectation operator, and $E[|\hat{N}|^2 \mid Z]$ is the expectation of noise spectrum N. In the following, the speech DFT coefficients are assumed to be Gaussian distributed and, therefore, the conditional probability density function (pdf) $f_{Z|N}(z|n)$ can be written [38] as

$$f_{Z|N}(z|n) = \frac{2z}{\lambda_x^2} \; e^{\left(-\frac{n^2+z^2}{\lambda_x^2}\right)} \; I_0\left(\frac{2nz}{\lambda_x^2}\right). \tag{3.8}$$

The $I_0(.)$ represents the 0^{th}-order modified Bessel function of the first kind and

λ_x^2 is clean speech variance. Inserting Eqs. (3.3) and (3.8) into (3.7), we get

$$E[|\hat{N}|^2 \ |Z] = \frac{\int_0^\infty n^{\nu+1} \ e^{(-\frac{n^2}{\lambda_x^2}-\frac{z^2}{\lambda_x^2}-n\beta)} I_0\Big(\frac{2nz}{\lambda_x^2}\Big) dn}{\int_0^\infty n^{\nu-1} \ e^{(-\frac{n^2}{\lambda_x^2}-\frac{n^2}{\lambda_x^2}-n\beta)} I_0\Big(\frac{2nz}{\lambda_x^2}\Big) dn}. \tag{3.9}$$

The numerator and denominator integrals do not have their closed form solutions. However, the solution of (3.9) can be found by approximating the Bessel function [114], i.e., for a given arguments x, it can be approximated to

$$I_0(\mathsf{x}) \approx \frac{1}{\sqrt{2\pi \mathsf{x}}} e^{\mathsf{x}}. \tag{3.10}$$

Clearly, from Figures 3.5(a) and (b), the Bessel function can be better approximated by argument ($\mathsf{x} \gtrsim 0.20$). Although for ($\mathsf{x} \lesssim 0.22$), the noise estimator in modulation domain shows insensitivity and, therefore, allow us to use this approximation. The closed form of both integrals can be derived by substituting

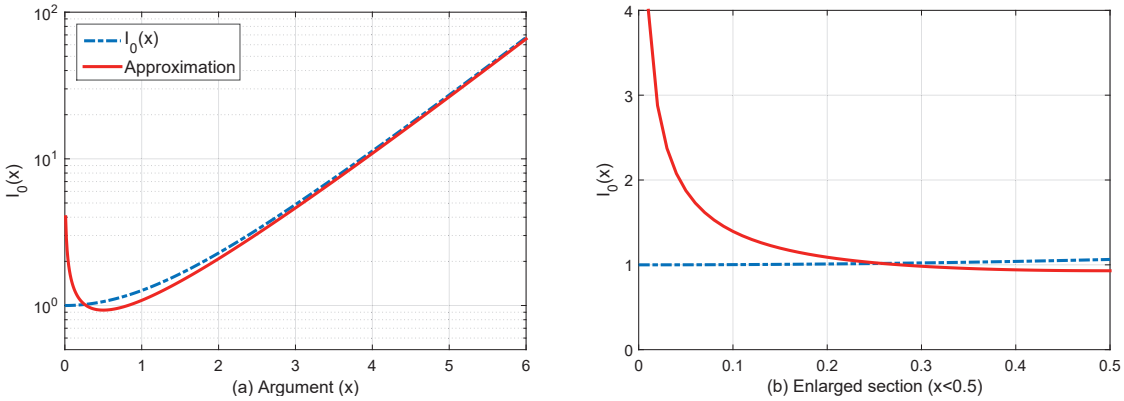


Figure 3.5: Bessel function of 0^{th} order and it's approximation from Eq. (3.10) for (a) argument x and (b) an enlarged section for x<0.50.

Eq. (3.10) into (3.9), we get

$$E[|\hat{N}|^2 \ |Z] = \frac{\int_0^\infty n^{\nu+\frac{1}{2}} \ e^{(-\frac{n^2}{\lambda_x^2}-n\beta+\frac{2nz}{\lambda_x^2})} dn}{\int_0^\infty n^{\nu-\frac{3}{2}} \ e^{(-\frac{n^2}{\lambda_x^2}-n\beta+\frac{2nz}{\lambda_x^2})} dn}. \tag{3.11}$$

Note that, second moment of the Gamma density provides the relationship between β and ν, i.e., $\beta^2 = \frac{\lambda_n^2}{\nu}$. Inserting β and using [115], solution of (3.11) is

obtained in terms of parabolic cylinder function $D_{\mathsf{n}}(\Phi)$ [116] of order n, i.e.,

$$E[|\hat{N}|^2 \ |Z] = \left(\frac{\xi}{2\gamma}\right)\frac{\Gamma(\nu+\frac{3}{2})}{\Gamma(\nu-\frac{1}{2})}\ \frac{D_{-(\nu+\frac{3}{2})}(\Phi)}{D_{-(\nu-\frac{1}{2})}(\Phi)}|Z|^2, \nu > \frac{1}{2} \tag{3.12}$$

where, $\Phi=\left(\sqrt{\nu\xi} - \sqrt{\frac{2\gamma}{\xi}}\right)$, and the $\Gamma(\mathsf{x})$ is the Gamma function defined only for positive events ($\mathsf{x} \geq 0$). The ξ and γ are the *a priori* and *a posteriori* SNRs and can be calculated by using $\xi=\frac{\lambda_x^2}{\lambda_n^2}$, and $\gamma=\frac{|Z^2|}{\lambda_n^2}$, respectively.

Given the noisy speech power $|Z^2(\tau,k,m)|$, and the *a priori* SNR estimate $\hat{\xi}(\tau,k,m)$, the noise power spectrum $|\hat{N}^2(\tau,k,m)|$, in the modulation domain can therefore be achieved by

$$|\hat{N}^2(\tau,k,m)| = \left(\frac{\xi(\tau,k,m)}{2\gamma(\tau,k,m)}\right)\frac{\Gamma(\nu+\frac{3}{2})}{\Gamma(\nu-\frac{1}{2})}\frac{D_{-(\nu+\frac{3}{2})}\Big[\Phi(\tau,k,m)\Big]}{D_{-(\nu-\frac{1}{2})}\Big[\Phi(\tau,k,m)\Big]}|Z^2(\tau,k,m)|. \tag{3.13}$$

From the investigative experiments conducted in subsection 3.2.2, it is empirically found that by using ν=2.60, the Gamma density closely approximates the real world noise DFT coefficients in both the acoustic and modulation domains. The importance of gamma density function is largely due to its relation to exponential and normal density functions [117]. However, the degree of asymmetry of the Gamma density diminishes by increasing the shape parameter ν. In other words, for larger ν, Gamma density approaches to the standard normal distribution.

From Fig. 3.6, it is noted that the strong noise spectral components (i.e., low instantaneous SNR) will be attenuated by almost the same amount for both MMSE [66] and the proposed noise methods. However the proposed gain value slowly decreases as the instantaneous SNR becomes higher. This particular characteristic of the proposed method is very useful for tracking the noise variations in the modulation domain as the spectral peakedness is relatively low in the modulation domain.

On the other hand, Eq. (3.13) depends on the *a priori* SNR, $\xi(\tau,k,m)=\frac{\lambda_x^2(\tau,k,m)}{\lambda_n^2(\tau,k,m)}$, and the *a posteriori* SNR $\gamma(\tau,k,m)=\frac{|Z^2(\tau,k,m)|}{\lambda_n^2(\tau,k,m)}$. For $\xi(\tau,k,m)$ estimate, the decision-directed approach [20] is used, as given by

$$\hat{\xi}(\tau,k,m) = \alpha\frac{|\hat{\mathcal{X}}(\tau-1,k,m)|^2}{|\hat{\lambda}_n^2(\tau,k,m)|} + (1-\alpha)\Big[\gamma(\tau,k,m)-1\Big], \tag{3.14}$$

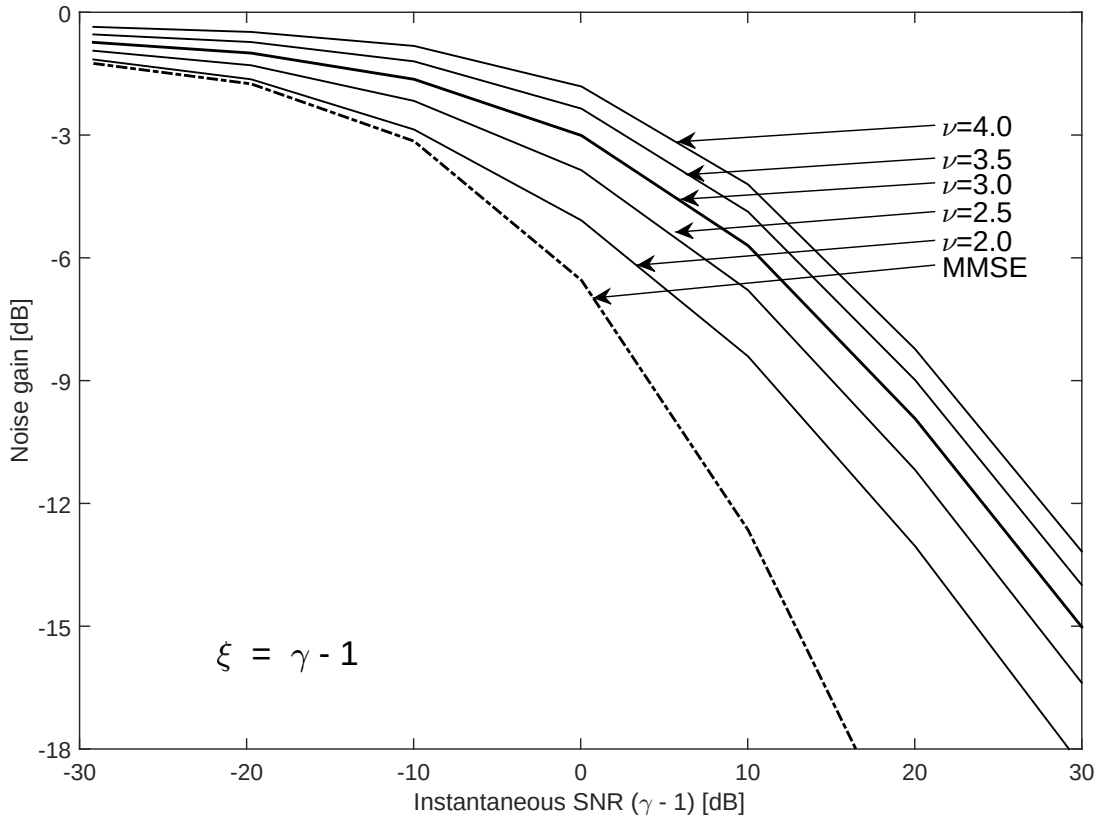


Figure 3.6: Plot of the proposed noise gain response for different scale parameter ν by considering the case of ξ equals to the instantaneous SNR (γ-1).

where,

$$\alpha_t = exp\big(\frac{-2.2R}{t_\tau f_\tau}\big), \tag{3.15}$$

is the smoothing factor with respect to time [118]. This approach is a linear combination of an estimate of the previous *a priori* SNR $\frac{|\hat{\mathcal{X}}(\tau-1,k,m)|^2}{|\hat{\lambda}^2_{mn}(\tau,k,m)|}$, and the maximum- likelihood (ML) SNR estimate $\Big[\gamma(\tau,k,m)-1\Big]$. Using past knowledge of the *a priori* SNR, the estimation process is influenced considerably for larger variances. This can be controlled by providing the time-based smoothing factor α for controlling the trade-off between speech distortion and random fluctuations [118]. If it is close to unity, a highly smoothed version of the ML estimate is achieved, but the drawback of reducing the variance (α close to unity) of *a priori* SNR estimate is that it can not respond quickly for sudden changes in the instantaneous SNR $\gamma(\tau,k,m)$ and that often leads to the transient distortion [119].

However, noise spectral coefficients are assumed to be more stationary than speech, it allows us to assume that the previous frame of the noise variance

$\hat{\lambda^2_{mn}}(\tau-1,k,m)$ is highly correlated with the present frame $\hat{\lambda^2_{mn}}(\tau,k,m)$, and therefore the noise variance estimate can be achieved by

$$\hat{\lambda}^2_{mn}(\tau,k,m) = \eta\ \hat{\lambda}^2_{mn}(\tau-1,k,m) + (1-\eta)\ |\hat{N}^2(\tau,k,m)|. \tag{3.16}$$

The $\hat{\lambda}^2_{mn}((\tau,k,m))$ is the modulation domain based noise variance estimate for current modulation time frame τ. Since the noise estimator (3.12) has to rely on the *a priori* SNR estimate, the noise estimation process entirely depends on the prior knowledge of the *a priori* SNR. The large estimation error usually occurs when there is a sudden change in *a priori* SNR due to speech onset where the estimated *a priori* SNR lags behind the actual *a priori* SNR. Note that the acoustic domain has relatively heavy tails in spectrum which may cause the large estimation error. To examine the behavior of proposed noise estimator in acoustic domain various investigative experiments have been conducted and results are compared in the following section 3.4.

3.4 Estimator's Performance in the Frequency Domain

The tracking capability and experimental results of the proposed frequency domain noise estimation against other conventional methods has been addressed in this section. For instance, to examine the tracking capability of the estimator towards a sudden changes in spectral power, highly non-stationary babble noise has been considered as it has the speech-like characteristics in the spectral domain. For this, two 6s long clean speech signals (one male and one female speaker) are taken from TIMIT database [44]. These clean speech signals are added together in order to get long enough to accommodate different noisy situations so that the estimator can show the capability to adopt the spectral change for varying input SNR. Moreover, to incorporate sudden change in noise statistics, there are two different levels of input SNR, i.e., 0dB and 20dB are considered. The 12s clean speech signal is degraded by using 20dB input SNR up to 3s and repeated from 6 to 9s. For a sudden increase in noise condition, 0dB input SNR is added from 3

to 6s and further from 9 to 12s. The clean speech signal with two levels of noise degraded speech (0dB and 20dB) which is considered for the analysis of different noise estimators are plotted Fig. 3.7.

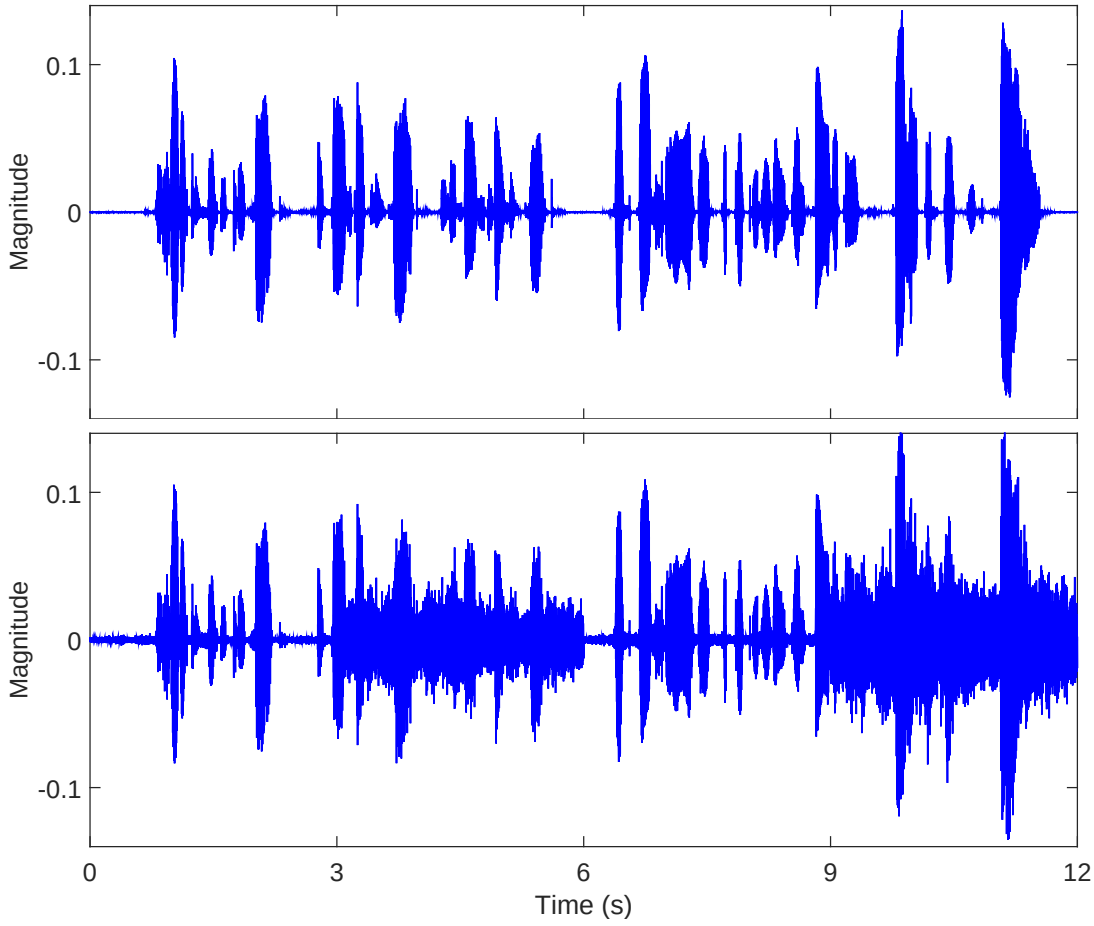


Figure 3.7: Plot of clean speech waveform (top) and clean speech segments 0-3sec, 6-9sec are degraded at 20dB input SNR, whilst segments 3-6sec and 9-12sec are degraded at 0dB input SNR by using non-stationary babble noise.

3.4.1 Drawbacks of MS and IMCRA Methods

As discussed in subsection 2.3.3, the noise estimation by MS method has relatively good accuracy than the spectral minima tracking [59] method. However, the method suffers by tracking capability as changes in noise estimate is delayed for any sudden noise spectral changes. This is because each window frame is divided into two sub-windows for increasing the estimation accuracy but, it requires larger memory size for processing to the window frame. Note that, the window length must be large enough to include the peaks of speech activity and short enough

to follow sudden noise variations. Therefore, reducing the frame size in MS method the delay problems arises because the time taken for processing each frame increases (almost double) largely.

As shown in Fig. 3.8, it is clear that the response of the MS estimator is insensitive especially when a sudden change in noise statistics occurs. It is noticed that for a sudden increase in true noise power at t>3s and t>9s, estimator fails to react and this results a delayed version of the noise estimate.

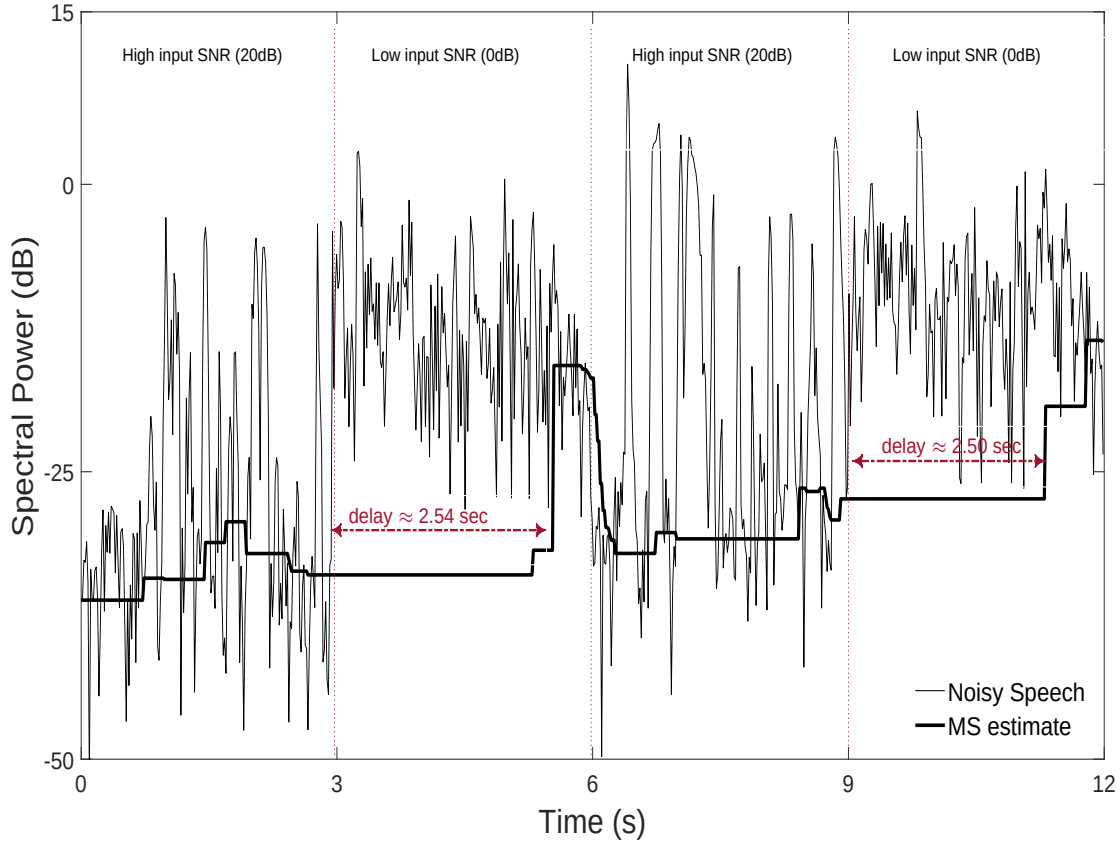


Figure 3.8: Plot of the noisy speech power with MS estimate by using two input SNR conditions (0dB & 20dB) of babble noise at a given frequency 500Hz.

Importantly, the estimator relies on the recursively updated noisy power given in Eq. (2.17), which means if the spectral minima for a given frame is unchanged, the estimator will fail to update the noise which causes the wrong estimation of the noise spectrum. This happens especially during low input SNR conditions (0dB time segments). Although the estimator has satisfactory results at high input SNRs (6-9s), a long delay (≈2.54s) to update the noise statistics is noticed when input SNR is low (0dB) which is clearly seen from Fig. 3.8. This behavior of the estimator is further validated from experimental results in terms of speech

intelligibility and SNRseg as plotted in Figs. 3.14 to 3.17 which indicated that the MS method do not provide much improvements in the frequency domain.

The recursive averaging based algorithms [62, 63] on the other hand estimate the noise spectral power recursively by using the speech presence probability (SPP). The SPP is obtained by comparing the ratio of the noisy power spectrum to its local minimum against a fixed threshold [62]. However, the spectral adaptation of noise using fixed threshold lags behind, especially when the noise increases abruptly. Also, the processing time to update the local minimum $\lambda^2_{min}(l,k)$ from a temporary estimate $\lambda^2_{tm}(l,k)$ increases as the temporary estimate updated for a given fixed window length as given in Eq. (2.22). For low input SNR conditions ($\leq$5dB), it is difficult to identify the speech presence frame where both the speech and noise signals almost same similar spectral energy levels. To address this problem [64] provides two frequency based fixed thresholds (2 and 5) but results remain same as the method uses same MCRA [62] principle. To up-

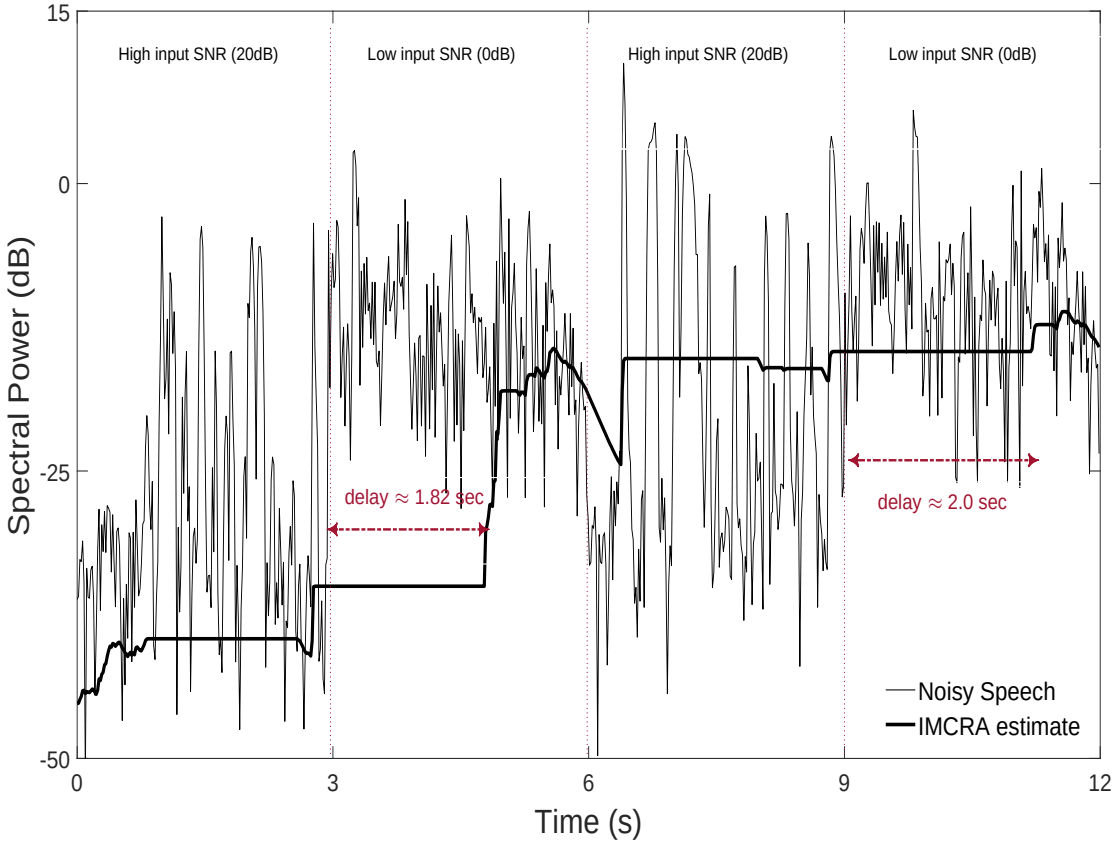


Figure 3.9: Plot of noisy speech power with IMCRA estimate by using two input SNR conditions (0dB & 20dB) of babble noise at a given frequency 500Hz.

date the noise power accurately, the SPP is replaced by the use of *a posteriori* probability in In [63]. Although, this method yields an improvement in noise tracking over [62], the computation of SPP is controlled by the minima values of a smoothed noisy power spectrum and therefore the noise estimate is influenced by tracking the minima of the spectrum. The problem arises, especially when the noise signal is non-stationary and has equal or higher energy than the speech signal. Additionally, the improved version of MCRA uses the principle of Minimum Statistics rule for tracking the minima of the spectrum, and therefore this IMCRA method faces similar problems of the estimation delay and processing time as compared to MS method [60]. Similar to MS method plotted in 3.8, the IMCRA updates the noise estimate with similar delay as shown in Fig. 3.9.

From the experimental results plotted in Figs. 3.14 to 3.17, both the MS and IMCRA methods do not have considerable improvements in both the speech intelligibility and segmental SNR measures as compared to the other considered methods. As noticed from Figs. 3.14a-3.17a, the STOI score even failed to improve compared to the unprocessed STOI observations. The reason for the reduction of the intelligibility in both methods is attributed to the fact that in MS method each window is divided into sub-window whilst the window length must be large enough to include the peaks of speech activity. Although dividing window in to sub-window may assist in improving segmental SNR measure as given in Figs. 3.14c-3.17c but simultaneously small window size introduces the distortion in speech activity and wrong noise estimation occurs which results in speech distortion. Since, the IMCRA estimator uses the MS principle, both MS and IMCRA methods therefore results in similar estimator's performance in terms of the loss of speech intelligibility.

Since, the MS and IMCRA methods provide comparably low performance mainly due to the delayed version of noise estimates, following subsections consider only the methods having better noise estimates relatively and promise better results in terms of intelligibility and segmental SNR improvements.

3.4.2 Comparison with MMSE-BC and soft-SPP

The Bayesian motivated MMSE based noise estimation algorithm is proposed in [65] where the noise spectral power is estimated by minimizing the mean square error (MSE) between the true and estimated noise spectrum by using the knowledge of the *a priori* SNR estimate. Since MS and IMCRA methods suffers from large estimation delay, i.e., ≈2.54s delay in MS method and ≈2.0s delay in IMCRA method that is clear from Figs. 3.8 and 3.9, the MMSE-BC method achieves good noise tracking comparatively. However, this method requires a biasing compensation factor. As such the *a priori* SNR estimate is delayed by at least one frame. Moreover, whenever the speech power is equal to noise power the estimator assumes that the frame has noise power only and the speech signal is misinterpreted by noise. The reason is that, when noise power is nearly equal to speech power, the *a priori* SNR estimate approaches to zero which is the case for noise-only frames and the estimated noise power increases while speech is present.

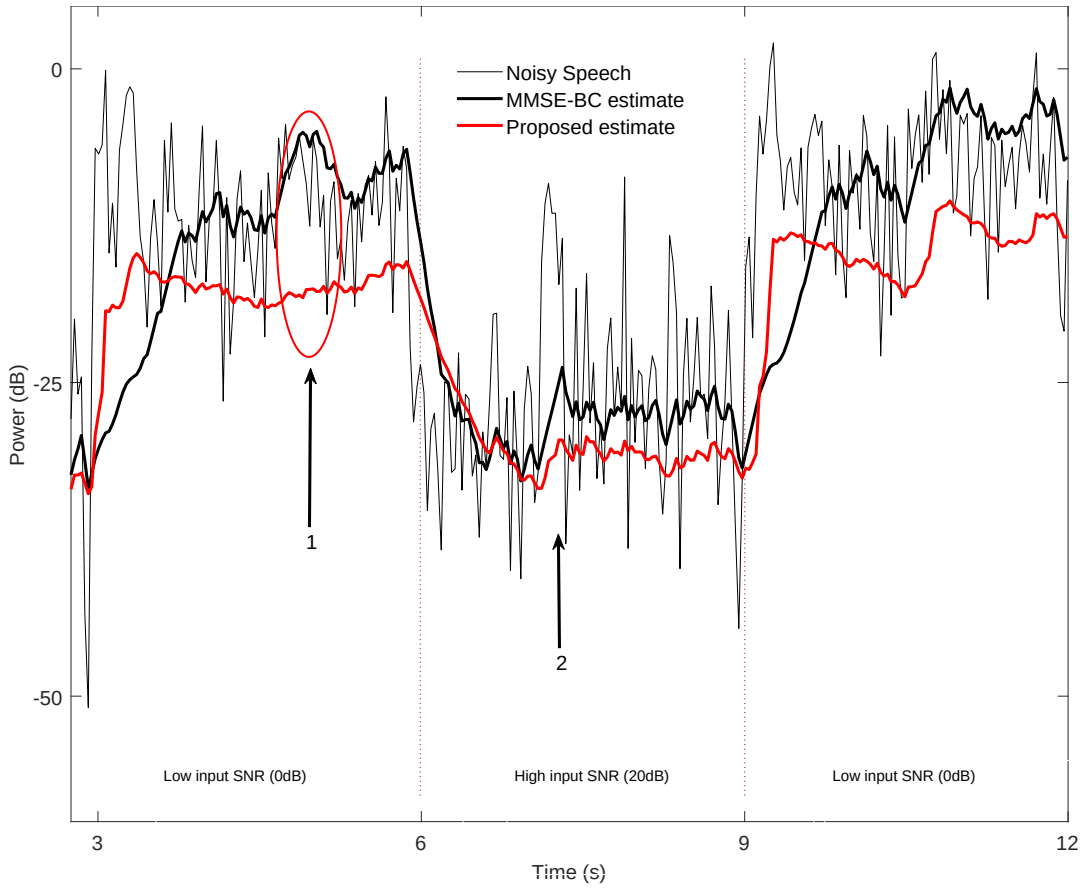


Figure 3.10: Plot of the noisy speech power and noise power spectrum estimated by MMSE-BC and proposed (3.16) noise methods at frequency 250Hz.

This noise over-estimation for a highly noisy condition (low input SNR 0dB) can be clearly seen from Fig. 3.10 where the MMSE-BC noise estimator clearly miss-interpreted speech as a noise and that results in a leakage of the speech spectral power in to the noise. Additionally, the biasing compensation factor

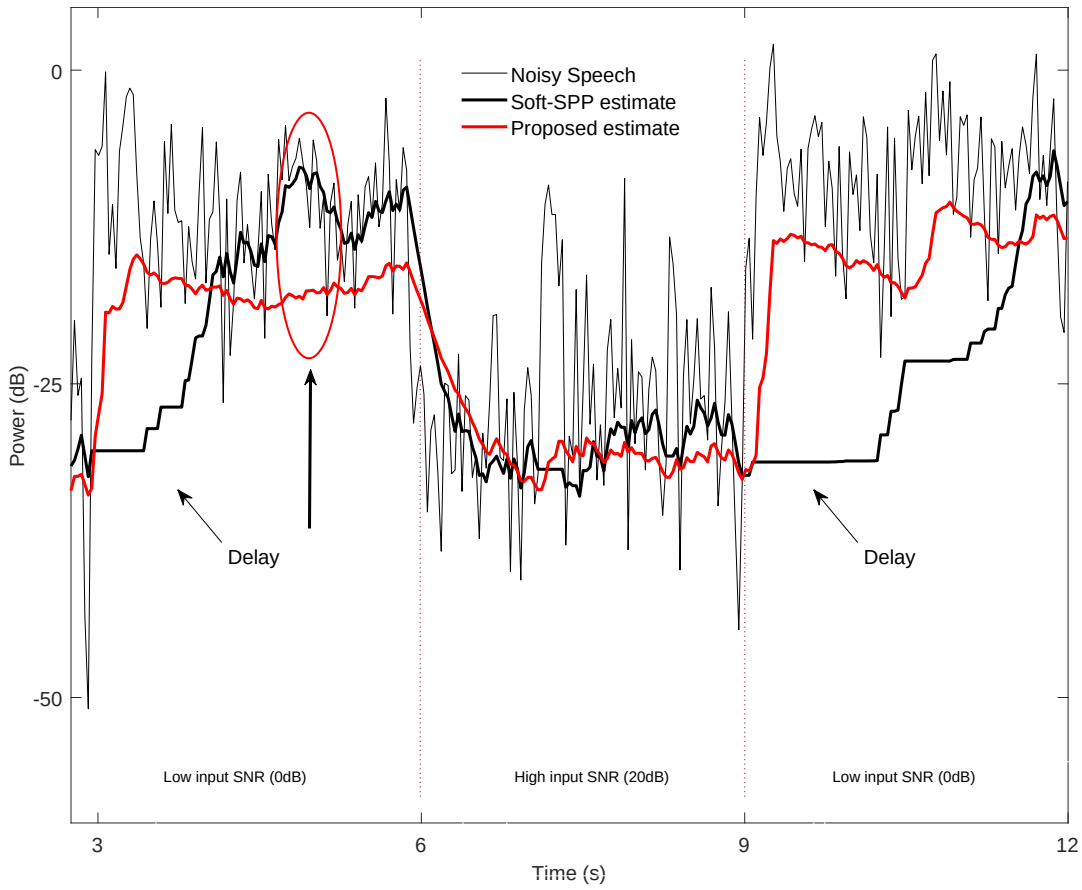


Figure 3.11: Plot of the noisy speech power and noise power spectrum estimated by Soft-SPP and proposed (3.16) noise methods at frequency 250Hz.

reacts only when the noise is under-estimated, especially for the estimated *a priori* SNR less than or equal to 10 dB as shown in Fig. 2.6. In other words, for a highly noisy conditions where both speech and noise signals have almost similar power, there is a high probability of erroneous noise estimation (over-estimation) that reduces the speech intelligibility.

Besides that, the proposed noise method adopts any sudden spectral change irrespective of the input SNR conditions. For example, when the noise power suddenly increases after t>3s and t> 9s, the proposed method reacts accordingly which gives a more methodical noise tracking. The reason of the better noise tracking may be that the Gamma density represents all time-varying noise signals

more closely compared to the aforementioned Gaussian density. Note that, the MMSE-BC method fails to be biased even by using a bias compensation factor especially for low input SNRs whilst without biasing compensation the proposed method successfully holds the leakage of the speech power to noise as indicated in Fig. 3.10. This may be the reason why proposed method achieves overall better tracking of highly non-stationary babble noise compared to the other conventional methods.

On the other hand, compensating the biasing effect in MMSE-BC [65], a soft-SPP with a fixed *a priori* SNR (ξ=15dB) is introduced in [66]. In this method, the Gaussian density function is assumed to derive to analytical model for all time-varying noise signals. Although replacing the biasing compensation factor by a fixed ξ reduces the computational complexity, it suffers by similar problem of over-estimation in the low input SNR conditions. As indicated in Fig. 3.11, the soft-SPP estimator deteriorates the speech spectral coefficients especially when both the speech and noise signals have same spectral power. As the soft-SPP estimator hold the Gaussian assumption for noise, estimator's complexity has been reduced greatly, but fixing the *a priori* SNR estimate to 15dB in [66] may restrict the capability to adopt fast spectral changes and that results in noise estimation delay as shown in Fig. 3.11. Both the MMSE-BC and soft-SPP estimators have almost similar performance in terms of the intelligibility and SNR improvements. The reason may be that these methods follow the same Gaussian assumption in the analytical derivation of the noise estimator.

3.4.3 Comparison with SIG Method

Recently, by using a sigmoid function to derive the soft-SPP has been proposed in [67]. The advantage of using the sigmoid based noise method is that, it provides the flexibility to use a range of the fixed *a priori* SNR (ξ_{H1}) from 12dB to 18dB, whilst the Gaussian assumption for time-varying noise signal suggests only 15dB for better noise power estimate [66]. The effect of using fixed ξ can be noticed from Fig. 3.11 where the soft-SPP method suffers by similar delay problem compared to the MMSE-BC method. However, Fig. 3.12 clearly indicates that the SIG method [67] that suggests a range of fixed *a priori* SNR (12dB, 18dB) has similar

delay problem in tracking the non-stationary noise signal. Although, sigmoid function may be an alternative to the Gaussian density assumption for noise, it fails in tracking a sudden changes in the input SNR levels. As shown in Figs. 3.14b to 3.17b the SIG method has lowest PESQ and SNRseg improvements compared to other methods. The proposed noise method on the other hand, provides better noise tracking in the frequency domain. This applies also for highly non-stationary babble noise without depending on the frequency as shown in Figs. 3.12 and 3.13 at 250Hz and 1kHz respectively. This individualistic behavior of the noise estimator results in a large improvement in both the intelligibility and segmental SNR of the frequency domain processed speech over the other conventional methods. More importantly, the proposed estimator adapts any sudden spectral change of the noise spectrum irrespective of the speech presence frame and stationarity nature of the noise signals. The reason may be that the proposed noise method uses the Gamma density as it provides minimum spectral

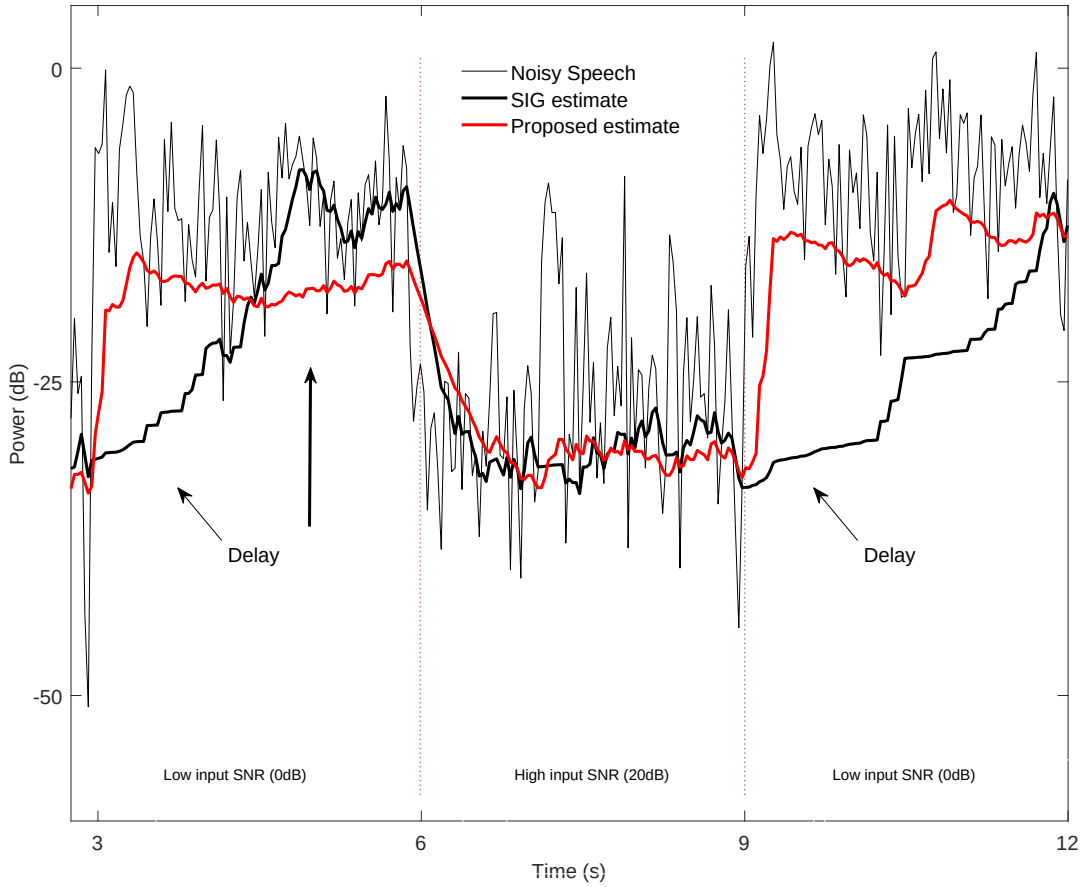


Figure 3.12: Plot of the noisy speech power and noise power estimated by SIG and proposed (3.16) noise methods at frequency 250Hz.

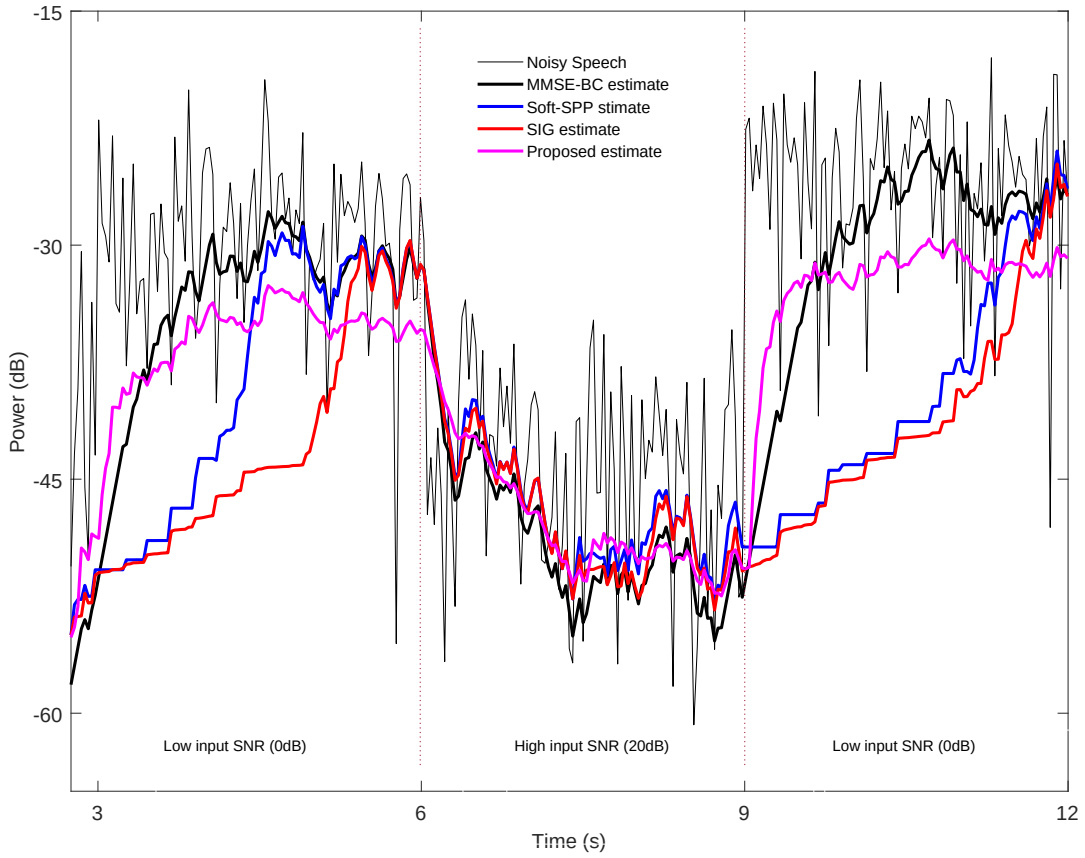
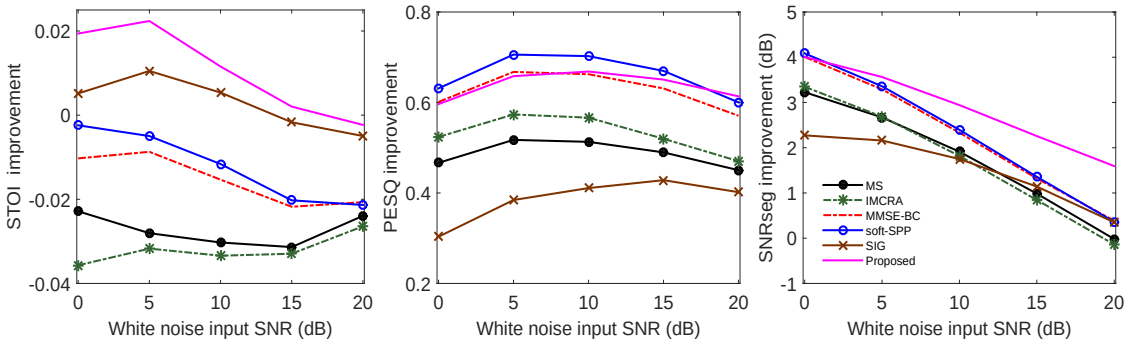


Figure 3.13: Plots of noisy speech power spectrum with the noise estimated by MMSE-BC, soft-SPP, SIG and proposed (3.16) noise methods at frequency 1kHz.

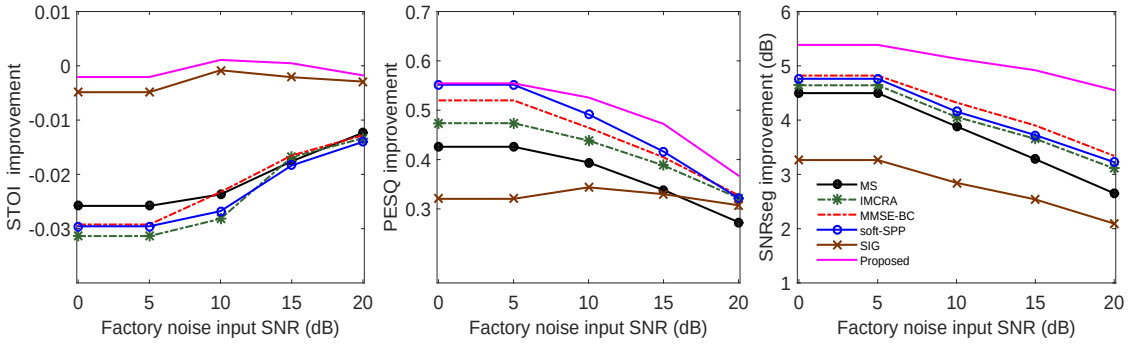
deviation from true spectrum of all time-varying noise signals.

Additionally, by using 250Hz frequency Fig. 3.13 plots the noise estimated by above methods [65–67] where it is noticed that for high input SNR (20dB) these methods provide almost similar results in tracking the actual noise spectrum. For a sudden increase change in noise level, i.e., t>3s and t>9s MMSE-BC method achieves faster noise update whilst both the soft-SPP and the SIG methods have almost similar delay. Based on these noise estimation methods, the experimental results in terms of the performance improvement are shown in Figs. 3.14, 3.15, 3.16, and 3.17 for stationary white noise, factory, street, and non-stationary babble noise, respectively.



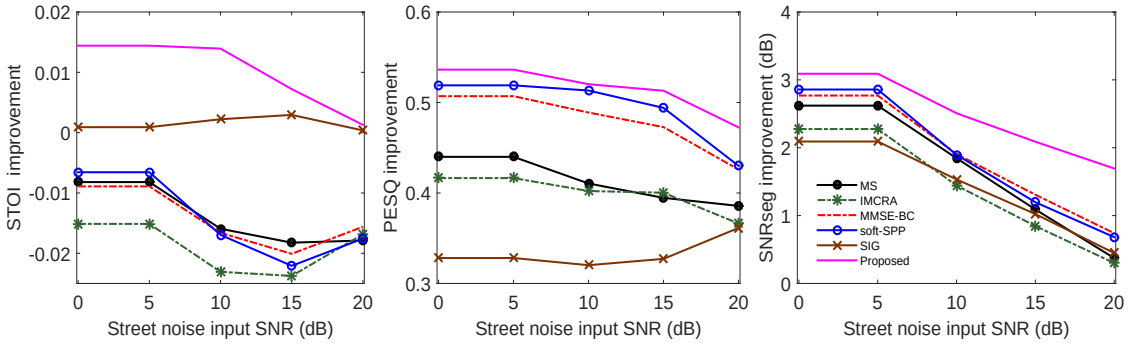
(a) STOI (b) PESQ (c) SNRseg

Figure 3.14: The frequency domain based mean STOI, PESQ, and SNRseg improvements for enhanced speech degraded by stationary white noise.



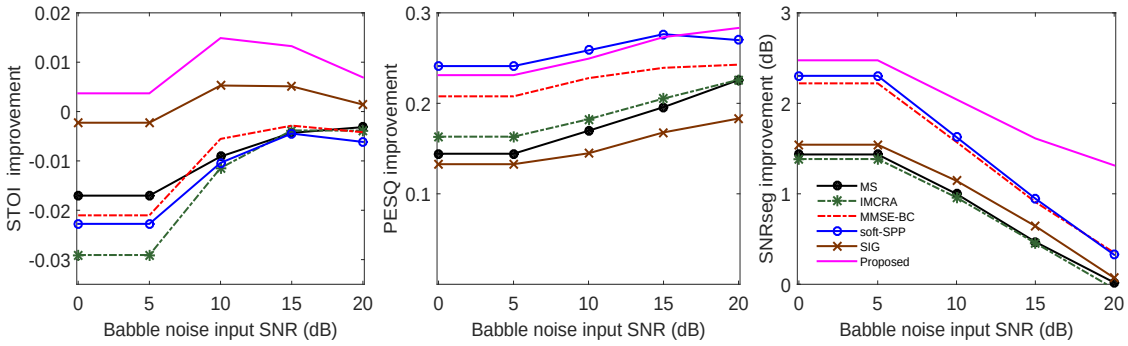
(a) STOI (b) PESQ (c) SNRseg

Figure 3.15: The frequency domain based mean STOI, PESQ, and SNRseg improvements for enhanced speech degraded by factory noise.



(a) STOI (b) PESQ (c) SNRseg

Figure 3.16: The frequency domain based mean STOI, PESQ, and SNRseg improvements for enhanced speech degraded by heavy street noise.



(a) STOI (b) PESQ (c) SNRseg

Figure 3.17: The frequency domain based mean STOI, PESQ, and SNRseg improvements for enhanced speech degraded by non-stationary babble noise.

3.5 Modulation Domain Noise Estimation

3.5.1 Experimental Settings

For primary AMS framework, the noisy speech is segmented by using a 512-point FFT length with a 6.25% frame shift (using Hamming window). By using this, each acoustic frequency bin has a bandwidth of 15.625 Hz and the individual acoustic envelope is sampled at 62.5 Hz. The modulation envelope is achieved by applying the modulation transform (i.e., secondary AMS framework) to each acoustic frequency bin by using the 32-point FFT achieving a 50% modulation frame shift (section 3.5.3). The decision directed approach (3.14) is used to estimate the *a priori* SNR with the smoothing factor α=0.98, and the enhanced stimuli is constructed by applying the estimated noise magnitude spectrum to the MMSE spectral amplitude estimation method with speech presence uncertainty (MMSE-SPU) [20].

3.5.2 Effect of Bias Compensation Factor

We have derived the Bayesian motivated noise estimator (3.12) that relies on *a priori* SNR ξ estimate. Since, the *a priori* SNR estimate depends on the previous frame of clean speech (3.14), the expectation of noise $E[\hat{N}|Z]$ is therefore biased and it requires a bias compensation factor. An estimator can be unbiased only when there is a perfect knowledge of the *a priori* SNR. To compensate for

this, an analytically derived bias compensation factor is proposed in [65], as

$$\left[\hat{\xi}(\tau,k,m)\right] = \frac{1}{\left(\left[1+\hat{\xi}(\tau,k,m)\right]\gamma_{ig}\left(\frac{1}{\hat{\xi}(\tau,k,m)+1},2\right)+e^{-\frac{1}{\hat{\xi}(\tau,k,m)+1}}\right)}, \tag{3.17}$$

where, $\gamma_{ig}(\rho,r)$ is the incomplete gamma function of ρ. Figure 3.18 plots the effect of the bias compensation factor as the SNR varies in the modulation domain. It

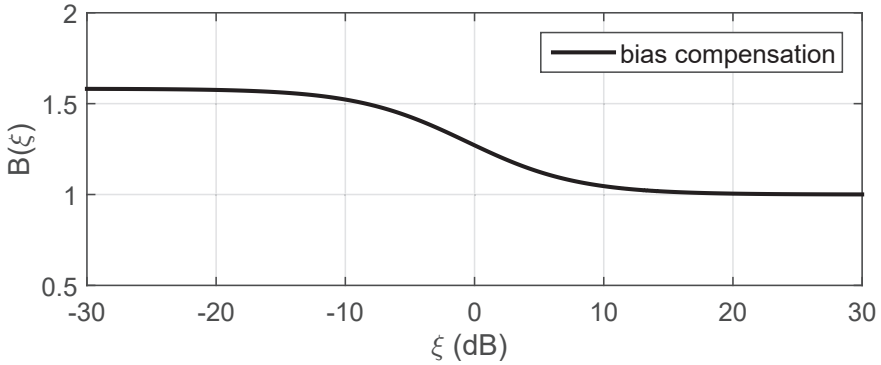


Figure 3.18: The bias compensation factor response (Eq. 3.17) with respect to the *a priori* SNR ξ.

is clear that the biasing factor $B\left[\hat{\xi}(\tau,k,m)\right]$ reacts only when the noise is under-estimated as $B\left[\hat{\xi}(\tau,k,m)\right] \geq 1$, especially for $\xi(\tau,k,m)$ less than or equal to 10 dB, whilst the noise estimator is considered to be unbiased for over-estimation. The noise power spectral density is then obtained by multiplying bias compensation factor $B\left[\hat{\xi}(\tau,k,m)\right]$ with $|\hat{N}^2(\tau,k,m)|$, as

$$\hat{\sigma}_n^2(\tau,k,m) = \begin{cases} \hat{N}^2(\tau,k,m)\ B\left[\hat{\xi}(\tau,k,m)\right] & \text{biased,} \\ \hat{N}^2(\tau,k,m) & \text{unbiased.} \end{cases} \tag{3.18}$$

Since noise is assumed to be more stationary than speech, the previous frame of the noise variance $|\hat{\lambda_n^2}(\tau-1,k,m)|$ is highly correlated with the present frame $|\hat{\lambda_n^2}(\tau,k,m)|$ and, the noise variance estimate can be obtained by a recursive averaging process, as given by

$$|\hat{\lambda}_n^2(\tau,k,m)| = \eta\ |\hat{\lambda}_n^2(\tau-1,k,m)| + (1-\eta)\ \hat{\sigma}_n^2(\tau,k,m), \tag{3.19}$$

where, η is the smoothing parameter. Clearly, a fast noise spectral change will cause a larger estimation error and, the noise estimator needs a bias compensation

factor. However, the bias compensation factor provides only a marginal improvements when the noise is under-estimated in the acoustic domain [65]. Table 3.1 shows that the modulation transform provides more stationarity compared to the acoustic domain. Due to slow varying spectral coefficients, noise signal will be better coupled with a distribution model. Moreover, the performance of the estimator will be improved by providing the correct tracking of any change in the noise statistics in the modulation domain. Therefore the noise estimator in the modulation domain may not need any bias compensation factor.

Biasing Effect in Acoustic Domain

To investigate the effect of bias compensation factor in the acoustic domain, investigative experiments are conducted by using both the biased and unbiased proposed noise estimators given in (3.18) and performance of the proposed noise estimator is compared with MMSE-BC [65]. From table 3.2, it is noted that the biased MMSE-BC [65] noise method in acoustic domain provides higher PESQ scores as compared to the unbiased MMSE-BC, which clearly indicates that the estimator strongly depends on the bias compensation factor in all the noisy situations.

On the other hand, proposed method uses the Gamma density (closely fitted) for the noise DFT coefficients in acoustic domain and consequently accurate noise estimate is achieved by tracking the noise spectral variations correctly.

From table 3.2, it is noted that the biased MMSE-BC [65] noise method in acoustic domain provides higher PESQ scores as compared to the unbiased MMSE-BC, which clearly indicates that the estimator strongly depends on the bias compensation factor that is irrespective of the types of noise and SNR levels. The reason may be that the estimator assumes Gaussian density to track the noise spectral variations. On the other hand, the proposed method uses the Gamma density (closely fitted) for the noise DFT coefficients in acoustic domain. As a result, the proposed method provides accurate noise estimate by tracking the noise spectral variations correctly. As we can see from table 3.2, the estimator has no bias compensation effect and provides similar intelligibility score irrespective of the noise types and SNR levels.

Table 3.2: The PESQ scores comparison in acoustic domain by using with and without bias compensation factor in both MMSE-BC and proposed methods.

Input SNR	White Noise				Factory Noise			
	MMSE-BC [65]		Proposed		MMSE-BC [65]		Proposed	
	Unbiased	Biased	Unbiased	Biased	Unbiased	Biased	Unbiased	Biased
0	1.9991	2.0165	2.0232	2.0258	2.6451	2.6876	2.7507	2.7573
5	2.3702	2.3996	2.4127	2.4167	3.0012	3.0323	3.0956	3.1030
10	2.7072	2.7536	2.7832	2.7934	3.2746	3.3136	3.3856	3.4050
15	3.0326	3.0767	3.1143	3.1273	3.5441	3.5919	3.6502	3.6727
20	3.3180	3.3769	3.4185	3.4359	3.7910	3.8388	3.8718	3.8895
Input SNR	Babble Noise				Street Noise			
	MMSE-BC [65]		Proposed		MMSE-BC [65]		Proposed	
	Unbiased	Biased	Unbiased	Biased	Unbiased	Biased	Unbiased	Biased
0	2.1057	2.1219	2.1390	2.1561	2.2657	2.2908	2.3645	2.3729
5	2.4569	2.4752	2.4958	2.5110	2.6046	2.6179	2.6945	2.7037
10	2.8108	2.8310	2.8516	2.8587	2.9372	2.9434	3.0136	3.0257
15	3.1532	3.1724	3.2024	3.2079	3.2384	3.2566	3.3229	3.3424
20	3.4816	3.5030	3.5303	3.5383	3.5032	3.5375	3.5946	3.6125

Biasing Effect in Modulation Domain

To substantiate the role of bias compensation factor in the modulation domain, performance of the proposed noise estimator has been analyzed and results are tabulated in Table 3.3. It is worth noting that both biased and unbiased proposed noise estimators provide similar performance in the modulation domain irrespective of the types of the noise and input SNR levels. In other words, the bias compensation factor has no influence on the proposed noise estimator even for low input SNR conditions. Recall that, the modulation domain provides comparably more predictable trend in the kurtosis measure as shown in table 3.1 which reduces the probability of miss-detection (wrong-estimate) due to the heavy tails in spectral changes. Due to this, the modulation domain seems to be insensitive to bias compensation factor irrespective to the noise types and different SNR conditions. Informal listening tests also reveal that the intelligibility of the enhanced speech by using the bias compensation factor based proposed noise method is similar to the unbiased method in the modulation domain. Additionally, the intelligibility of the enhanced speech is improved in the modulation domain as

compared to the acoustic domain. and therefore the estimator is considered to be unbiased in the modulation domain.

Table 3.3: Mean PESQ scores for the Modulation domain based proposed noise method using with and without bias compensation factor.

Input SNR	White Noise		Factory Noise		Babble Noise		Street Noise	
	Unbiased	Biased	Unbiased	Biased	Unbiased	Biased	Unbiased	Biased
0	2.0098	2.0142	2.7798	2.7834	2.1428	2.1467	2.3544	2.3517
5	2.4146	2.4205	3.1191	3.1226	2.5102	2.5163	2.6849	2.6832
10	2.8036	2.8096	3.4393	3.4432	2.8702	2.8758	3.0242	3.0258
15	3.1362	3.1407	3.7146	3.7143	3.2301	3.2323	3.3481	3.3494
20	3.4494	3.4524	3.9226	3.9230	3.5607	3.5625	3.6236	3.6259

Informal listening tests also reveal that the intelligibility of the enhanced speech by using the bias compensation factor based proposed noise method is similar to the unbiased method in the modulation domain as well as for increasing the input SNR conditions, the intelligibility of the enhanced speech is improved in the modulation domain compared to the acoustic domain.

3.5.3 Role of FFT Size and Shift in Modulation Domain

The literature on the best FFT size and shift in the implementation of the modulation domain remains inconclusive. Different settings have been reported with mixed performance [24, 37, 120–122]. As smaller FFT size provides higher intelligibility [27], whilst speech based applications such as hearing aid devices prefer smaller FFT size. Applications such as speech coding, on the other hand, prefer better quality over intelligibility [46] and, the selection of modulation FFT size may differ by having a larger FFT size. Besides, small acoustic frame shift provides larger time frames that allow estimator to be more adaptive and thus providing improved intelligibility. However, that translates to increased number of sampling points and it often leads to a considerable amount of time to process the modulation frames.

Therefore to achieve the suitable FFT size and shift for acoustic domain, we have conducted comprehensive experiments by using various acoustic FFT lengths (64 to 1024) and varying frame shifts. The intelligibility based performance of the

enhanced stimuli is analyzed by using PESQ and STOI scores which are shown in Figs. 3.19 and 3.20. Fig. 3.19 shows that acoustic frame shift with 6.25% provides the overall best speech preservation of speech intelligibility Whilst by increasing the frame shift, the stationarity of the spectrum reduces and the estimator may fail to response quickly for sudden spectral changes, which would result decrease in speech intelligibility. Although the PESQ performance by using 12.5% is also acceptable and can be used, the STOI score in Fig. 3.20 clearly indicates that the 6.25% is better than other frame shift for all the noisy conditions. Informal listening tests also verify these differences of enhanced stimuli. On the other hand, both 256 and 512-point FFTs perform well in terms of PESQ score while STOI score strongly supports 512-point FFT rather than 256-point FFT. Further increasing FFT size to 1024 or even higher, the intelligibility of the enhanced speech is degraded. In this experiment it is found that the acoustic FFT size of 512 along with 6.25% frame shift yields overall best performance.

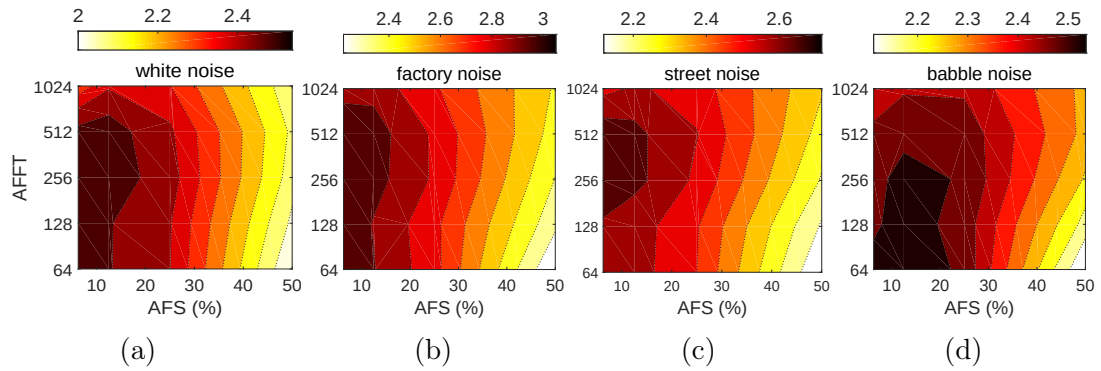


Figure 3.19: The mean PESQ score for varying AFFT and AFS by using a fixed 32-point MFFT achieving 50% MFS.

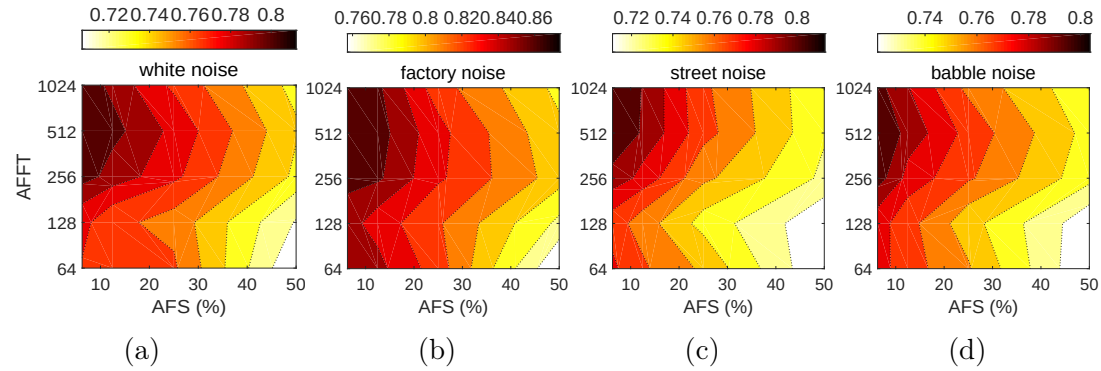


Figure 3.20: The mean STOI score for varying AFFT and AFS by using a fixed 32-point MFFT achieving 50% MFS.

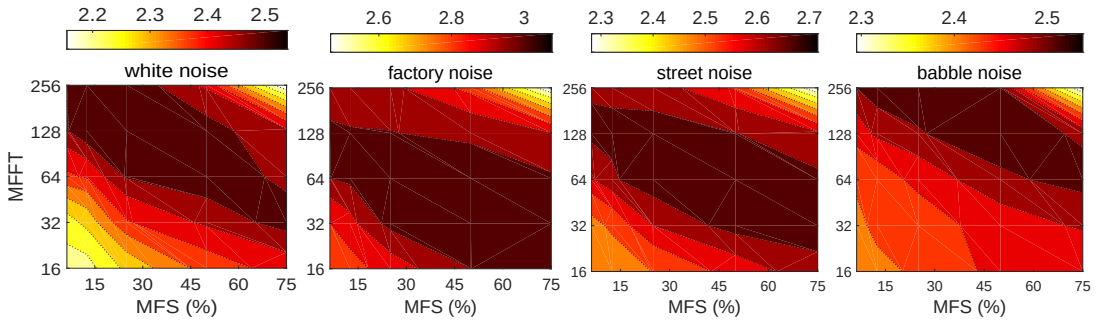


Figure 3.21: The mean PESQ score for varying MFFT and MFS by using a fixed 512-point AFFT achieving 6.25% AFS.

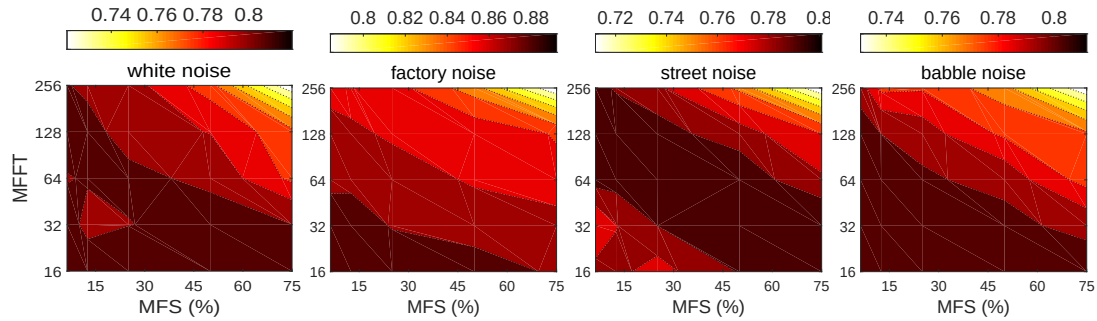


Figure 3.22: The mean STOI score for varying MFFT and MFS by using a fixed 512-point AFFT achieving 6.25% AFS.

Another experiment was performed find the suitable FFT size and shift for modulation transform by using different modulation FFT size and shifts and achieved results are shown in Figs. 3.21 and 3.22. Results indicate 32 and 64-point modulation FFTs achieve overall better intelligibility while increasing further to 256 or higher causes the reduction in speech intelligibility. It also shows that a 16-point modulation FFT would lead to degradation of the overall quality of the speech. As suggested in [27,37,103], the intelligibility of the enhanced speech can usually be improved by reducing the frame shifts. Contrary to this, results based by varying the modulation frame shift shown in Figs. 3.21 and 3.22, clearly, reveal that 50% modulation frame shift delivers higher PESQ and STOI scores at 64-point and 32-point FFT, respectively, for the case of white noise degraded speech. Similarly, the STOI scores for babble and street noise degraded speech, 32-point FFT gives higher intelligibility score. In this experiment, the 32-point modulation FFT size with 50% frame shifting achieves overall quality of the enhanced speech.

Based on these experimental findings, the frame size of 512-point FFT size with 6.25% shift is selected for processing of the acoustic magnitude spectrum whilst for modulation transform, a 32-point FFT size achieving a 50% shift is considered for the experiments presented in the next subsections.

3.5.4 Modulation Results and Discussions

In this section, modulation based performance of the proposed noise estimator is presented. The modulation domain based performance of the proposed noise estimator is compared with aforementioned noise estimation methods in terms of their intelligibility based STOI, and PESQ scores as well as the segmental SNR measures. Results for varying noise stationarity such as stationary white noise, factory noise, non-stationary babble noise and heavy street noise are shown in Figs. 3.23, 3.24, 3.26, and 3.25, respectively. The acoustic domain based performance of proposed noise estimator has been investigated and compared in the section 3.4, whilst the STOI based performance is tabled in Tables 3.4, 3.5, 3.6 and 3.7, by comparing the intelligibility of the enhanced speech processed through acoustic and modulation domain, respectively. For stationary white noise, it is clear from Fig. 3.23 that, the proposed noise estimator in modulation domain achieves better noise estimate and provides higher intelligibility scores (PESQ, STOI) with the segmental SNR improvement over the existing methods. As can be seen from Fig. 3.23b, the PESQ score is improving throughout, whilst the STOI improvement as given in Fig. 3.23a clearly shows that the proposed noise method successfully restore the speech signal with a large improvement in STOI score where other methods fail to preserve the originality (negative) of the enhanced speech. However for low input SNR ($\leq$ 10 dB), the MS and IMCRA methods do not react quickly to sudden spectral changes, which results in an under-estimation (residual) of the noise and loss of intelligibility is perceived. Unlike frequency domain, the SIG method however improves the performances by tracking the noise power in modulation domain, relatively. The MMSE-BC and MMSE-SPP methods, on the other hand, improve the noise tracking capability by providing better noise estimates as compared to the MS and IMCRA methods but as noticed from Fig. 3.23a, they fail to improve the STOI score at low

Table 3.4: mean STOI score for stationary white noise corrupted speech processed in acoustic and modulation domains.

Input	Acoustic domain							Modulation domain						
SNR	Noisy	MS	IMCRA	MMSE-BC	soft-SPP	SIG	**Proposed**	Noisy	MS	IMCRA	MMSE-BC	soft-SPP	SIG	**Proposed**
0	0.6655	0.6426	0.6298	0.6552	0.6630	0.6707	**0.6849**	0.6655	0.6512	0.6342	0.6677	0.6737	0.6738	**0.6924**
5	0.7807	0.7527	0.7490	0.7720	0.7758	0.7913	**0.8031**	0.7807	0.7661	0.7467	0.7734	0.7776	0.7765	**0.8058**
10	0.8752	0.8449	0.8418	0.8599	0.8636	0.8806	**0.8867**	0.8752	0.8544	0.8432	0.8599	0.8631	0.8588	**0.8883**
15	0.9386	0.9073	0.9057	0.9169	0.9184	0.9370	**0.9407**	0.9386	0.9140	0.9067	0.9172	0.9189	0.9133	**0.9404**
20	0.9736	0.9497	0.9472	0.9529	0.9523	0.9687	**0.9713**	0.9736	0.9508	0.9437	0.9497	0.9499	0.9461	**0.9693**

Table 3.5: mean STOI score for factory noise corrupted speech processed in acoustic and modulation domains.

Input	Acoustic domain							Modulation domain						
SNR	Noisy	MS	IMCRA	MMSE-BC	soft-SPP	SIG	**Proposed**	Noisy	MS	IMCRA	MMSE-BC	soft-SPP	SIG	**Proposed**
0	0.7933	0.7689	0.7639	0.7669	0.7704	0.7909	**0.7903**	0.7933	0.7611	0.7513	0.7611	0.7659	0.7643	**0.7885**
5	0.8701	0.8443	0.8387	0.8408	0.8405	0.8652	**0.8680**	0.8701	0.8419	0.8327	0.8387	0.8407	0.8409	**0.8662**
10	0.9273	0.9036	0.8991	0.9041	0.9006	0.9266	**0.9284**	0.9273	0.9088	0.9034	0.9007	0.8995	0.9002	**0.9278**
15	0.9648	0.9472	0.9480	0.9483	0.9465	0.9628	**0.9653**	0.9648	0.9491	0.9483	0.9482	0.9474	0.9447	**0.9674**
20	0.9849	0.9726	0.9716	0.9721	0.9709	0.9821	**0.9832**	0.9849	0.9754	0.9748	0.9750	0.9744	0.9701	**0.9855**

Table 3.6: mean STOI score for non-stationary babble noise corrupted speech processed in acoustic and modulation domains.

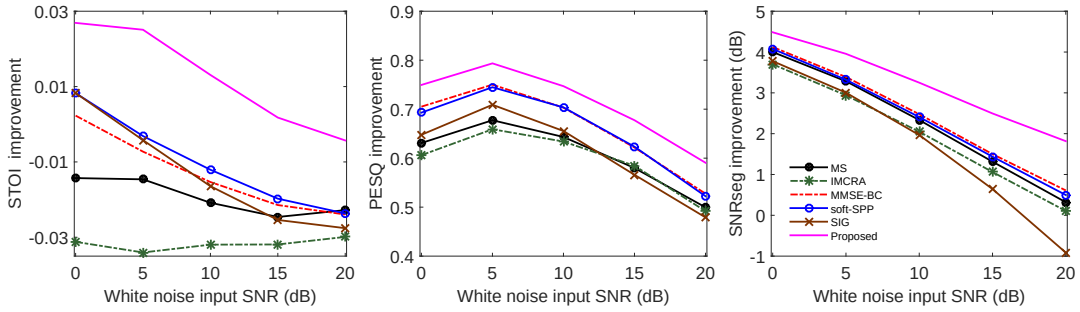
Input	Acoustic domain							Modulation domain						
SNR	Noisy	MS	IMCRA	MMSE-BC	soft-SPP	SIG	**Proposed**	Noisy	MS	IMCRA	MMSE-BC	soft-SPP	SIG	**Proposed**
0	0.7030	0.6727	0.6602	0.6603	0.6597	0.6876	**0.6844**	0.7030	0.6634	0.6568	0.6545	0.6560	0.6584	**0.6808**
5	0.7954	0.7783	0.7663	0.7744	0.7726	0.7932	**0.7991**	0.7954	0.7746	0.7651	0.7727	0.7731	0.7757	**0.7992**
10	0.8731	0.8640	0.8616	0.8675	0.8627	0.8784	**0.8880**	0.8731	0.8604	0.8540	0.8652	0.8624	0.8638	**0.8885**
15	0.9301	0.9258	0.9263	0.9272	0.9256	0.9353	**0.9433**	0.9301	0.9256	0.9206	0.9277	0.9267	0.9252	**0.9441**
20	0.9657	0.9626	0.9618	0.9615	0.9596	0.9672	**0.9726**	0.9657	0.9596	0.9584	0.9631	0.9619	0.9586	**0.9743**

Table 3.7: mean STOI score for street noise corrupted speech processed in acoustic and modulation domains

Input	Acoustic domain							Modulation domain						
SNR	Noisy	MS	IMCRA	MMSE-BC	soft-SPP	SIG	**Proposed**	Noisy	MS	IMCRA	MMSE-BC	soft-SPP	SIG	**Proposed**
0	0.6865	0.6774	0.6744	0.6807	0.6853	0.6887	**0.6969**	0.6865	0.6690	0.6594	0.6758	0.6814	0.6804	**0.6974**
5	0.7772	0.7690	0.7621	0.7683	0.7707	0.7782	**0.7916**	0.7772	0.7631	0.7532	0.7706	0.7740	0.7727	**0.7982**
10	0.8611	0.8452	0.8381	0.8445	0.8442	0.8634	**0.8751**	0.8611	0.8485	0.8344	0.8550	0.8562	0.8546	**0.8854**
15	0.9266	0.9083	0.9028	0.9065	0.9045	0.9296	**0.9338**	0.9266	0.9107	0.8985	0.9177	0.9168	0.9146	**0.9425**
20	0.9676	0.9497	0.9508	0.9520	0.9500	0.9680	**0.9689**	0.9676	0.9555	0.9490	0.9590	0.9575	0.9554	**0.9745**

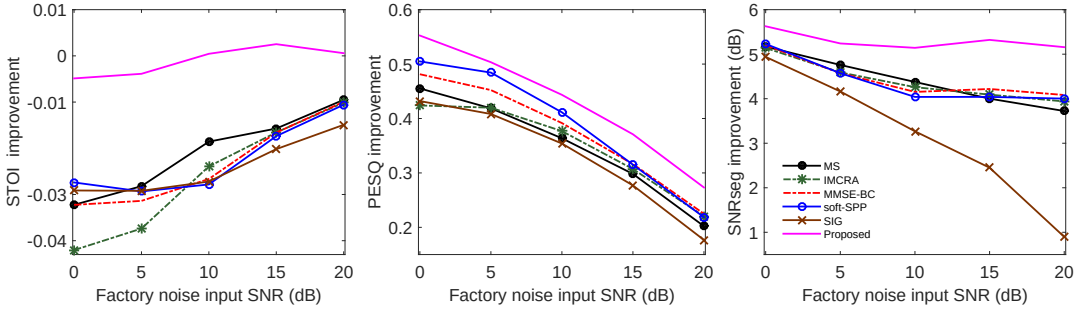
input SNR conditions. Recall in sub-section 3.2.2 the Gamma density models the modulation based stationary white noise histogram more closely (96%) than the Rayleigh density (93%). This is evident in the results as the proposed noise estimator achieves better noise tracking and, the performance is greatly improved over the aforementioned noise estimators. Similar observation is obtained for the segmental SNR improvement, in particular for the input SNR equal or greater than 5 dB, as shown in Fig. 3.23c. Besides that, Table 3.4 shows that the proposed method achieves improvements in terms of STOI score over other methods in both the acoustic and modulation domain, which supports the modulation domain as a better alternative for improving the speech intelligibility.

The performance of factory noise corrupted speech is shown in Figs. 3.24, and the STOI score of the corrupted speech processed through acoustic and modulation domains is presented in Table 3.5. Clearly, the PESQ and segmental SNR improvements of the noise estimators are relatively similar to white noise, while the STOI score in Fig. 3.24a, indicates that the proposed noise estimator is suc-



(a) STOI (b) PESQ (c) SNRseg

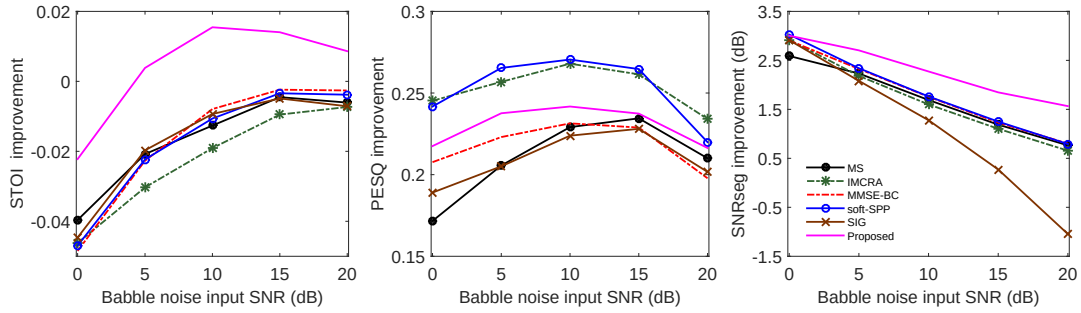
Figure 3.23: The modulation based performance in terms of (a) STOI, (b) PESQ, and (c) SNRseg improvements for stationary white noise degraded speech.



(a) STOI (b) PESQ (c) SNRseg

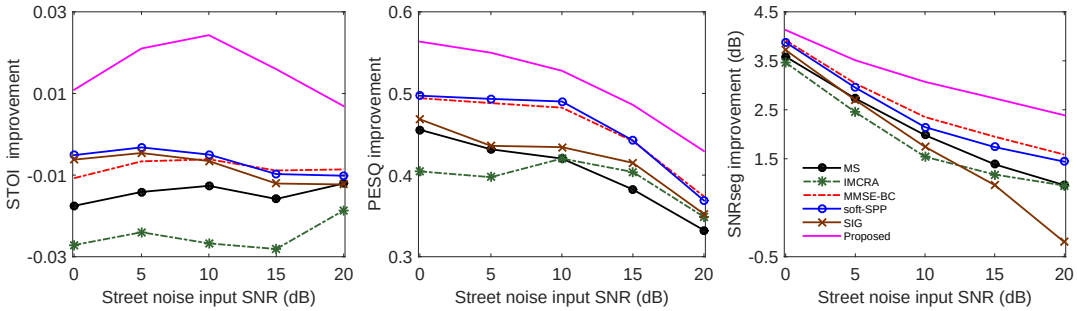
Figure 3.24: The modulation based performance in terms of (a) STOI, (b) PESQ, and (c) SNRseg improvements for factory noise degraded speech.

cessful in preserving the quality of the speech whilst reduction in original speech intelligibility score (negative) is observed for other aforementioned noise methods. In fact, the intelligibility follows that of the original unprocessed signal and yet noise is suppressed. Figs. 3.25, and Table 3.6 illustrate the performance of the



(a) STOI (b) PESQ (c) SNRseg

Figure 3.25: The modulation based performance in terms of (a) STOI, (b) PESQ, and (c) SNRseg improvements for heavy street noise degraded speech.



(a) STOI (b) PESQ (c) SNRseg

Figure 3.26: The modulation based performance in terms of (a) STOI, (b) PESQ, and (c) SNRseg improvements for non-stationary babble noise degraded speech.

speech degraded by non-stationary babble noise. As given, for low input SNR conditions ($\leq$ 5 dB), improving the speech intelligibility is a difficult task. This is because, babble has various peaks distributed randomly in the spectral domain and the estimator may erroneously mistake the noise as speech and results in a large estimation error.

Nevertheless, the proposed method outperforms the other conventional methods. As we can see in Fig. 3.25a, the intelligibility in terms of STOI score is reduced for input SNR less than or equal to 5 dB. For higher input SNR ($\geq$ 5 dB), the noisy speech processed by using the proposed noise estimator provides a notable improvement in speech intelligibility. Similar improvement is observed in terms of segmental SNR measure given in Fig. 3.25c, where the proposed

noise method provides considerable improvements even for poor input SNR. The performance of heavy street noise degraded speech in terms of the STOI, PESQ and SNRseg improvements are shown in Fig. 3.26, and the comparison between acoustic and modulation domains based STOI score is tabled in Table 3.7. From these results, the proposed noise estimator clearly outperforms the other methods in the different noise settings.

3.5.5 Subjective Evaluation and Discussion

Subjective listening test was conducted in the form of AB listening tests that determined parameter preference [37]. In each subjective test, listening tests were conducted in a quiet room. the participants were explained the procedure during a short practice session and they were free to listen the stimuli multiple times if required at a comfortable listening level. A computer based three label were given and participants were asked to make their subjective preference. The first and second options were used to indicate a preference for the corresponding stimuli, while the third option was used to indicate a similar preference for both stimuli.

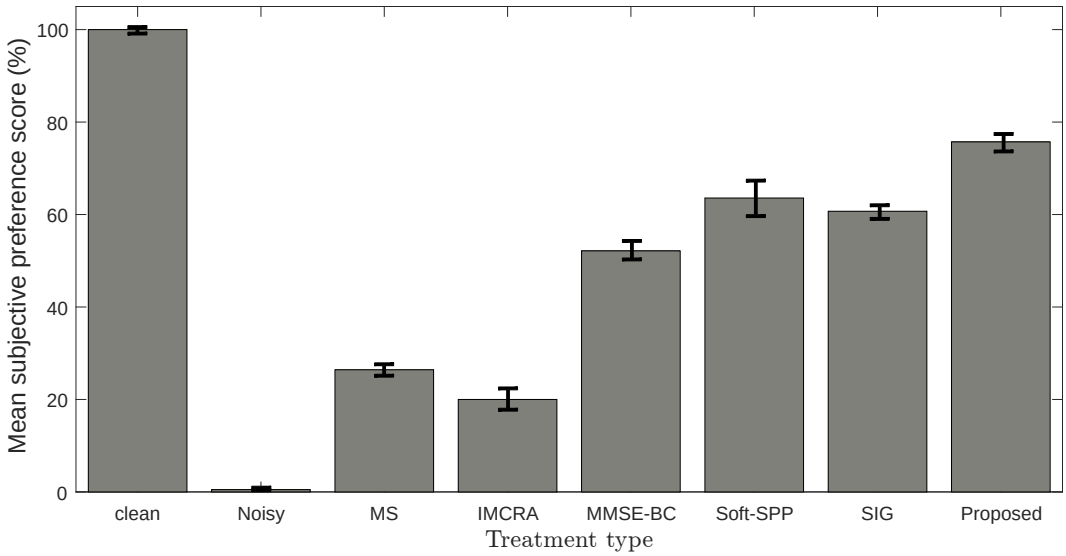


Figure 3.27: Mean subjective preference scores (%) with standard error bars for (a) clean; (b) noisy (degraded at 5 dB AWGN); and stimuli generated by using the following modulation domain based treatment types: (c) Minimum Statistics; (d) IMCRA; (e) MMSE-BC; (f) Soft-SPP; (g) SIG; and (h) Proposed (3.16) noise methods.

The listeners were instructed to use the third option only when they did not

prefer one stimulus over the other. Pair-wise scoring was used, with a score of +1 awarded to the preferred treatment, and 0 to the other. For the similar preference response, each treatment was awarded a score of +0.5. Two TIMIT sentences [44], of 5s long belonging to one male and one female speaker and degraded with 5 dB AWGN as well as babble noise were used. The complete test included 112 stimuli pairs for comparisons and total 12 listeners participated in the test.

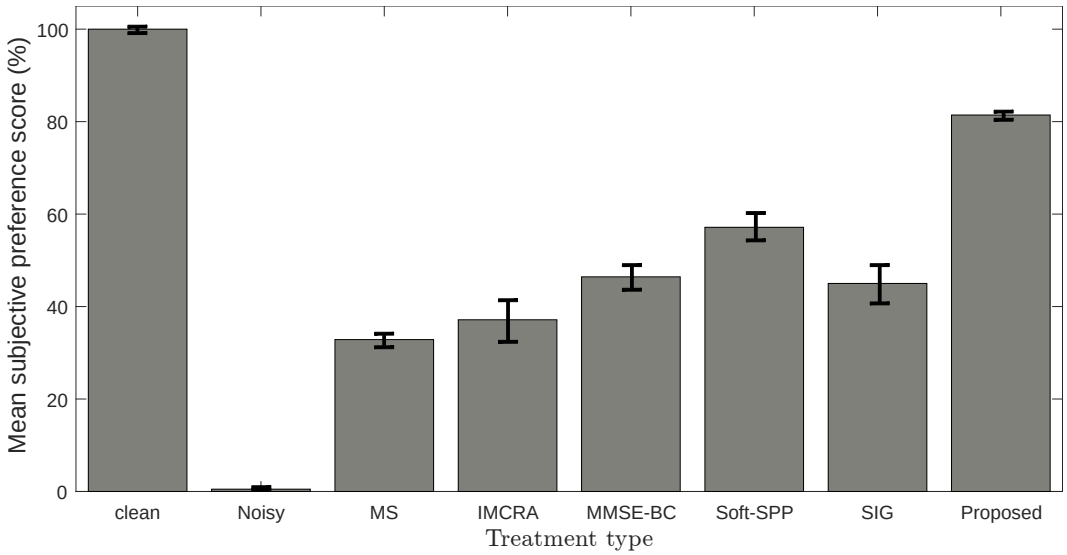


Figure 3.28: Mean subjective preference scores (%) with standard error bars for (a) clean; (b) noisy (degraded at 5 dB Babble noise); and stimuli generated by using the following modulation domain based treatment types: (c) Minimum Statistics; (d) IMCRA; (e) MMSE-BC; (f) Soft-SPP; (g) SIG; and (h) Proposed (3.16) noise methods.

The results from mean subjective preference scores for stmuli degraded by stationary white noise are shown in Fig. 3.27 whilst, Fig. 3.28 shows the mean subjective preference scores for stimuli degraded by babble noise. The proposed method has significantly higher score than the state of the art methods that can be clearly seen from Figs. 3.27 and 3.27 in which the listeners indicate their preference towards the proposed method. Noting that the performance of Soft-SPP [66] over the other methods is better and this can also be seen fron the objective measures shown in Figs. 3.23, 3.24, 3.25, and 3.26.

Therefore, it is worth noting that the proposed noise estimator in modulation domain achieves a constant performance in all varying noise conditions, and the quality improvement in objective as well as subjective scores is shown from the different measures over existing estimators.

3.6 Summary

This chapter presents a Bayesian motivated noise magnitude estimator in both the frequency and modulation domain. Both acoustic and modulation histograms are fitted with the Gaussian, Rayleigh, and Gamma densities. The K-S test shows that the Gamma density provides the minimum deviation from the noise histograms and found to be best-suited noise model for all varying stationarity of noise. The Bayesian motivated MMSE noise estimator is derived by using the Gamma density. Moreover, the kurtosis measure by using both the acoustic and modulation domain spectra reveal that the noisy speech spectrum in the acoustic domain has heavy tails as compared to the spectral variation in the modulation domain. Due to this nature of modulation spectrum, the leakage of speech power in to noise will be reduced and therefore the proposed noise estimator can adapt to spectral changes efficiently without requiring any bias compensation factor in the modulation domain.

As the selection of appropriate FFT size and frame shift achieves better intelligibility in the modulation domain, experiment was investigated for various combinations of FFT size and shift and it was found that the combination of 512-point AFFT with 6.25% AFS delivers optimum intelligibility for primary AMS framework. Whilst, for secondary AMS framework (modulation transform) 64-point MFFT achieving 50% MFS provides optimal quality of the enhanced speech. The experimental findings showed that the proposed noise estimator successfully restores the speech signal and provides measurable improvement in terms of PESQ, STOI, and segmental SNR in the modulation domain.

$\sim * \sim$

Chapter 4

Framework of Modulation Domain Bayesian Noise Estimators

Mind is true laboratory, where behind the illusions, we uncover the laws of truth.

–Sir Jagadish Chandra Bose.

4.1 Introduction

Results from chapter 3 clearly revealed that the contribution of modulation domain towards speech intelligibility improvements is significantly large and the modulation domain has been registered to be a better alternative to time-frequency domain for speech enhancement, as speech intelligibility is closely linked with the modulation spectrum. Additionally, Schimmel in his dissertation [28] reports that the energy from two different signals (e.g., speech and noise signals) in the modulation domain is largely non-overlapping. This is also supported by psychoacoustic research, which indicates that the human auditorial system segregates sound in the modulation domain [29]. This suggests greater demarcation between the noise and the speech in the modulation domain. Also, various research, which dates back to the early 90s, indicate that modulation domain processing often results in higher intelligibility of speech [27, 30–34]. Moreover, the intelligible components of the speech signal are mostly confined to the modulation frequency band of 1Hz to 16Hz and therefore processing can be made to concentrate on the

relevant bands [35, 36].

Reducing the background noise level on the other hand will invariably produce some speech distortions and a trade-off between the desired noise reduction and the undesired speech distortions must be achieved. In brief, the performance of a noise estimator is determined by the spectral information available in the Fourier domain. Based on these informations, the noise estimator can provide a good trade-off between the speech distortion (noise over-estimation) and the residual noise. However, to control the trade-off, different noise methods carry-out estimation process by considering the properties of noise DFT coefficients independently. For example, the Mean Square Error (MSE) cost function provides same result for both positive and negative estimation errors. Frankly speaking, perceptually the positive error (i.e., the estimated magnitude is smaller than the true magnitude) and negative error (i.e., the estimated magnitude is larger than the true magnitude) are not the same in speech enhancement based applications. Moreover, by considering the non-stationary noise signal the estimation process is affected due to sudden variation in the noise statistics. Due to this, these methods work satisfactory for the stationary noise signals and failed to track the correct information of spectral variation of non-stationary noise like babble or street for noise signals.

The problem of sudden change in noise statistics has already been addressed in chapter 3, where it is clearly shown that the existing methods fail to adopt a sudden noise spectral change. This is may be one of the reason why the current state of the art noise estimators do not contribute much in the improvement of the speech intelligibility. Various other factors that are also responsible for the absence of intelligibility improvement with existing conventional noise methods. The majority of these factors center around the fact that none of the existing algorithms are designed to improve speech intelligibility, as they utilize a cost function that does not necessarily correlate with speech intelligibility [81]. The common approach is to perform statistical estimation by minimizing the Bayes Risk of the squared-error of the spectral amplitude cost function using the Bayesian approach, which leads to the subsequent and traditional Minimum Mean-Square Error (MMSE) short-time spectral amplitude (STSA) estimator [20]. The MMSE

method, however, pays no attention to positive or negative differences between the true and estimated spectra, and the signal will be attenuated for a positive difference while a negative difference provides the spectral amplification. These two perceptual differences in minimizing the mean squared error (MMSE) cost function cannot be assumed to be equivalent in terms of speech intelligibility. This is because an improvement in SNR does not mean an increased in intelligibility. However, the effectiveness of the Bayesian-based methods is well accepted for single-channel speech enhancement, as various methods modify the traditional cost function to achieve more subjectively meaningful speech estimators. In [21], several perceptually-motivated spectral amplitude cost functions have been derived by stressing more on the spectral valleys rather than spectral peaks (formants). The reason indirectly provided in support of the spectral valleys is that it is associated with auditory masking effects. Particularly, Weighted Euclidean (WE) and Weighted COSH (WCOSH) cost functions as these cost functions control the estimator by providing a weighting function. The estimator based on these cost functions gives the best performance, especially for diminishing the residual noise effect and producing overall better speech quality.

By effective use of the modulation domain findings from chapter 3, this chapter derives the family of Bayesian noise estimators by generalizing the WE and COSH cost functions.

4.2 Bayesian Theory

For single-channel speech enhancement, the most fundamental techniques are closely connected to the Bayesian methods, as minimizing a Byes risk for a given cost function achieves a variety of estimators. In fact, the maximum a posteriori (MAP) estimator, minimum mean square error (MMSE) and Maximum likelihood (ML) estimators can be derived from the different Bayes risk cost functions. The Bayesian estimators based on perceptually motivated cost functions in place of traditional cost function are closely related to a Bayes risk [20, 21, 38, 123]. As the single-channel speech enhancement has limited information, the Bayesian estimation plays a prominent role in reducing the complication and in estimating

the most accurate spectral coefficients [20]. This is because the Bayesian theory minimizes the Baye's risk function, which includes a posterior probability model of the unknown parameters (given from the observation vector) and a cost function. As mentioned previously, the posterior probability density function (pdf) depends on how relatively the noise pdf is peaked, i.e., the likelihood pdf depends on the posterior pdf. The more peaked the noise pdf, the larger the estimation error will be, and as a result, the greater the influence on the outcome of the estimation process. Conversely, a uniform pdf will have no influence on the estimation [43].

As detailed in section 3.3, the central components in Bayesian estimation process are the posterior pdf $f_{N|Z}(n|z)$, and the cost function $C(n, \hat{n})$. The posterior pdf $f_{N|Z}(n|z)$ of the noise signal N, for a given noisy signal Z, is given as

$$f_{N|Z}(n|z) = \frac{f_{Z|N}(z|n)\ f_N(n)}{f_Z(z)}. \tag{4.1}$$

For a given noisy observation, $f_Z(z)$ is a constant [43] and has only a normalizing effect. Therefore, the behavior of $f_{N|Z}(n|z)$ depends only on two variables, i.e.,

$$f_{N|Z}(n|z) \propto f_{Z|N}(z|n)\ f_N(n), \tag{4.2}$$

where $f_{Z|N}(z|n)$ is the likelihood that the observation signal z is generated by the noise vector n and $f_N(n)$ is the prior pdf of the noise vector n. As stated earlier, the relative influence of the likelihood pdf $f_{Z|N}(z|n)$ and the prior pdf $f_N(n)$ on the posterior pdf $f_{N|Z}(n|z)$ depends on the shape of these two probability functions. In other words, the more peaked a pdf, the more it will influence the outcome of the estimation process [43]. Secondly, the estimation accuracy of an estimator depends on the behavior of the cost function as it represents the difference between true and estimated random variables and, therefore, it is often possible to correlate the cost function with the estimation error, as

$$\mathcal{E}_n = n - \mathbb{E}[\hat{n}], \quad \text{and} \quad C(n, \hat{n}) = \mathcal{E}_n. \tag{4.3}$$

For minimizing the estimation error $\mathcal{E}_n$, $C(n, \hat{n})$ allows tuning by providing the spectral weight so that the estimator can achieve the desired outcomes.

The Bayesian estimation of a parameter vector n is based on the minimization of a Bayesian risk function defined as

$$\mathcal{R}(\hat{n}, z) = \mathbb{E}[C(n, \hat{n})]. \tag{4.4a}$$

$$= \int_n \int_z C(n, \hat{n}) f_{Z,N}(z, n)\ dz\ dn. \tag{4.4b}$$

$$= \int_n \int_z C(n, \hat{n}) f_{N|Z}(n|z) f_Z(z)\ dz\ dn. \tag{4.4c}$$

Since, $f_Z(z)$ has no effect on the risk-minimization process, Eq. (4.4c) can be simplified in terms of the conditional risk function $f_{N|Z}(n|z)$, as follows

$$\mathcal{R}(\hat{n}, z) = \int_n \int_z C(n, \hat{n}) f_{N|Z}(n|z)\ dz\ dn. \tag{4.5}$$

The estimate of an unknown variable is obtained by assuming that the $\mathcal{R}(\hat{n}, z)$ is differentiable and has a well-defined minimum as

$$\mathbb{E}[\hat{N}|Z] = \arg\min_{\hat{n}} \Big[\mathcal{R}(\hat{n}, z)\Big]. \tag{4.6a}$$

$$= \arg\min_{\hat{n}} \left[\int_n \int_z C(n, \hat{n}) f_{N|Z}(n|z)\ dz\ dn\right]. \tag{4.6b}$$

From Eq. (4.2), the estimator can be derived by differentiating Eq. (4.6) and setting the gradient to zero as

$$\mathbb{E}[\hat{N}|Z] = \arg\operatorname{zero}_{\hat{n}} \left[\int_n \int_z C(n, \hat{n})\ f_{N|Z}(n|z)\ f_Z(z)\ dz\ dn\right]. \tag{4.7}$$

From Eq. (4.7), the nature of the derived estimator depends on the given cost function $C(n, \hat{n})$, the $f_{Z|N}(z|n)$, and most importantly, the distribution function $f_N(n)$ of that parameter to be estimated. As, the probability distribution function $f_N(n)$ which fits best for time-varying noise DFT coefficients in both frequency and modulation domain has already been explored in chapter 3, the following sections concentrate on the problem associated with a suitable cost function $C(n, \hat{n})$ to track the non-stationary noise DFT coefficients in the modulation domain. For simplicity, it is assumed that the speech and noise spectral coefficients have a Gaussian distributed and therefore the marginal probability density function of

both speech and noise DFT coefficients are given by

$$f_{Z,\theta_{mz}}(z) = \frac{z}{\sqrt{2\pi\lambda_{mz}^2}}\, e^{-\left[\frac{z^2}{\lambda_{mz}^2}\right]}, \tag{4.8}$$

$$f_{N,\theta_{mn}}(n) = \frac{n}{\sqrt{2\pi\lambda_{mn}^2}}\, e^{-\left[\frac{n^2}{\lambda_{mn}^2}\right]}. \tag{4.9}$$

The additivity and independence Gaussian assumption of speech and noise gives the joint probability distribution function $f_{Z|N,\Delta}(z|n,\theta_{mx})$ in spectral domain, as

$$f_{Z|N,\Delta}(z|n,\theta_{mx}) = \frac{1}{\sqrt{2\pi\lambda_{mx}^2}}\, e^{-\left[\frac{z^2+n^2}{\lambda_{mx}^2}\right]}\, I_0\left(\frac{2nz}{\lambda_{mx}^2}\right), \tag{4.10}$$

where, $I_0(.)$ represents the modified Bessel function of order zero, and θ_{mz}, θ_{mx}, and θ_{mn} are the given modulation phase spectrum of the noisy speech, clean speech and noise signal, respectively. To this end, we define

$$\frac{1}{\lambda_{mz}^2} = \frac{1}{\lambda_{mx}^2} + \frac{1}{\lambda_{mn}^2}, \tag{4.11}$$

$$\nu_k = \frac{\xi}{\xi+1}\gamma, \quad \nu_n = \frac{\nu_k}{\xi^2} \quad \text{and} \quad s = \nu_n\, \lambda_{mz}^2, \tag{4.12}$$

and the *a priori* and the *a posteriori* SNRs are given by

$$\xi = \frac{\lambda_{mx}^2}{\lambda_{mn}^2}, \quad \gamma = \frac{Z^2}{\lambda_{mn}^2}. \tag{4.13}$$

By using above marginal (Eq. 4.9) and joint (Eq. 4.10) probability distributions, the following section derives the perceptually motivated Bayesian noise estimators. Also, the behavior of these noise estimators towards the spectral changes has been compared with the speech estimators derived by using that distortion measure.

4.3 Distortion Measures

The statistically meaningful and more prominent cost functions for tracking the highly non-stationary speech DFT coefficients in the frequency domain have been suggested in [21]. Later in [124], it was found in a subjective comparison of many

different speech enhancement methods that the Bayesian approach performed in general better than the other methods in terms of the amount of speech distortion introduced by the processing and the background noise reduction in the frequency domain. On the other hand, experimental findings from table 3.1 in chapter 3 clearly indicate that the modulation domain has relatively low tails in terms of the spectral variation and DFT coefficients are comparably more stationary. Moreover, various physiological and psychoacoustic findings show modulation domain processing highly correlates with improvement in speech intelligibility. Therefore these cost function based Bayesian noise estimators may give more promising results by improving the speech quality and intelligibility in the short-time modulation domain.

4.3.1 Minimum Mean Square Error (MMSE) Measures

The traditional MMSE estimator minimize the mean square error (MSE) between true and estimated parameter by using the cost function as

$$C(n, \hat{n}) = (n - \hat{n})^2, \tag{4.14}$$

and the risk function using the square error cost function (Eq. 4.14) can be written as

$$\mathcal{R}(\hat{n}, z) = \int_n \int_z C(n, \hat{n})\ f_{N|Z}(n|z)\ dz\ dn. \tag{4.15}$$

Since, the probability functions $f_{N|Z}(n|z)$ are non-negative, minimizing the inner integral minimize the risk function. Differentiating the inner integral with respect to $\hat{n}$, yields

$$\frac{\partial}{\partial \hat{n}}\Big[\mathcal{R}(\hat{n}, z)\Big] = \frac{\partial}{\partial \hat{n}}\left[\int_n \int_z C(n, \hat{n})\ f_{N|Z}(n|z)\ dz\ dn\right] = 0. \tag{4.16}$$

$$= \frac{\partial}{\partial \hat{n}}\left[\int_n \int_z (n - \hat{n})^2\ f_{N|Z}(n|z)\ dz\ dn\right] = 0. \tag{4.17}$$

$$= -2\int_n (n - \hat{n}) f_{N|Z}(n|z)\ dn = 0. \tag{4.18}$$

Setting Eq. (4.18) to zero, the estimator that minimizes the mean square error is given by

$$\mathbb{E}[\hat{N}|Z] = \frac{\int_0^\infty n\ f_{N|Z}(n|z)\ dn}{\int_0^\infty f_{N|Z}(n|z)\ dn}. \tag{4.19}$$

Substituting Eq. (4.9) and Eq. (4.10) in Eq. (4.19), and neglecting constant terms, gives

$$\mathbb{E}[\hat{N}|Z] = \frac{\int_0^\infty n^2\ e^{-\frac{n^2}{\lambda^2}}\ I_0\left(\frac{2nz}{\lambda_{mx}^2}\right)\ dn}{\int_0^\infty n\ e^{-\frac{n^2}{\lambda^2}}\ I_0\left(\frac{2nz}{\lambda_{mx}^2}\right)\ dn}. \tag{4.20}$$

The solution of the integrals in Eq. (4.20) can be obtained from [115, 6.631] and, the MMSE based noise estimator is given by

$$\mathbb{E}[\hat{N}|Z] = \Gamma(\frac{3}{2})\ \sqrt{\frac{\xi}{\gamma(\xi+1)}}\ \Phi(-\frac{1}{2},1;-\nu_n)\ Z, \tag{4.21}$$

where the confluent hyper-geometric function

$$\Phi(-\frac{1}{2},1;-\nu_n) = e^{-\nu_n/2}\big[(1+\nu_n)I_0(\nu_n/2) + \nu_n\ I_1(\nu_n/2)\big], \tag{4.22}$$

is written in terms of the modified Bessel functions of order zero (I_0) and one (I_1) [125, A1.31a], whilst, $\Gamma(.)$ is the gamma function. Simplifying Eq. 4.21 by letting $\nu_n = \frac{\nu_k}{\xi^2}$ from 4.12, the MMSE noise gain (G_N), can be written as

$$G_N = \Gamma(\frac{3}{2})\ \frac{\sqrt{\nu_k}}{\gamma}\ e^{-\left(\frac{\nu_k}{2\xi^2}\right)}\left[\left(1+\frac{\nu_k}{\xi^2}\right)I_0\left(\frac{\nu_k}{2\xi^2}\right) + \left(\frac{\nu_k}{\xi^2}\right)\ I_1\left(\frac{\nu_k}{2\xi^2}\right)\right]. \tag{4.23}$$

Note that, the similar estimator has been derived for estimating the MMSE-short-time spectral amplitude (MMSE-STSA) in [20, Eq. 7], as

$$G_S = \Gamma(\frac{3}{2})\ \frac{\sqrt{\nu_k}}{\gamma}\ e^{-\left(\frac{\nu_k}{2}\right)}\left[\left(1+\nu_k\right)I_0\left(\frac{\nu_k}{2}\right) + \left(\nu_k\right)\ I_1\left(\frac{\nu_k}{2}\right)\right]. \tag{4.24}$$

Comparing both Eq. (4.23), and (4.24), it is clear that the behavior of the MMSE noise estimator is dependent upon the *a priori* SNR. Previous studies [39, 65, 66] suggest that estimating the noise in MMSE sense over speech may provide an alternative by reducing the background noise successfully and an improvement

in overall speech quality and speech intelligibility will therefore be achieved. To realize the effectiveness of the MMSE noise method in real-time, Fig. 4.2 plots the MMSE noise gain along with the Wiener noise gain response by using the instantaneous variation of both the *a priori* and *a posteriori* SNRs. Since the noise estimator uses the *a priori* SNR that is depending on the *a posteriori* SNR estimate for a given frame, here we consider the following relation to achieve a real-time variation of the gain function by using

$$\xi = \max\big[\gamma - 1, 0\big], \tag{4.25}$$

to estimate the *a priori* SNR for plotting the gain functions.

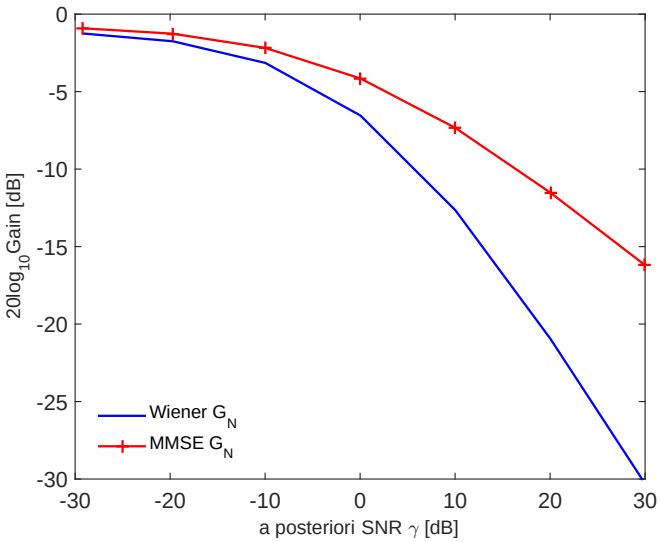


Figure 4.1: The plot of the Wiener noise gain and, the MMSE noise gain function derived in Eq. (4.24).

It can be observed from Fig. 4.1 that when γ decreases, the MMSE gain G_N gravitates towards the Wiener G_N whilst, for high SNR condition, the Wiener G_N has the smaller value relatively which translate the noise under-estimation and results in the higher residual noise. On the other hand, MMSE G_N has higher noise gain value that may provide the speech distortion due to noise over-estimation. The performance the MMSE based G_N has been detailed in chapter 5.

4.3.2 β-Order MMSE Measures

The MMSE based method [20, 22] provides better estimate as its solution is derived from mathematical derivation by minimizing the squared-error cost function based on the Gaussian model and statistical independence assumption. Although the elimination of musical noise by minimizing the mean squared error (MSE) estimator is effective [126,127], it might not be subjectively meaningful as small and large squared estimation errors might not necessarily correspond to good and poor speech quality respectively. To overcome the above problems and shortcomings of the squared-error cost function, β-Order MMSE estimator will have a much wider range of gain values and thus more flexible and effective for estimating the noise spectral components provided that the value can be appropriately adapted for different noise signal component strengths. Moreover in the given squared-error cost function, the parameter β used as exponent actually controls the associated estimator gain function and as a result the trade-off between speech distortion and noise reduction can be managed. By tuning the correct β value, the derived estimator provides the correlation the human auditory system indirectly similar to modulation domain. Although, a β-MMSE method for estimating the speech amplitudes has been derived in [127], it suggests to use only a positive value of the exponent for tracking the speech DFT amplitudes. The β-MMSE noise estimator can be derived by using the parameter β as an exponent to the noise random variable n given in Eq. (4.14), as

$$C(n,\hat{n}) = (n^{\beta} - \hat{n}^{\beta})^2, \tag{4.26}$$

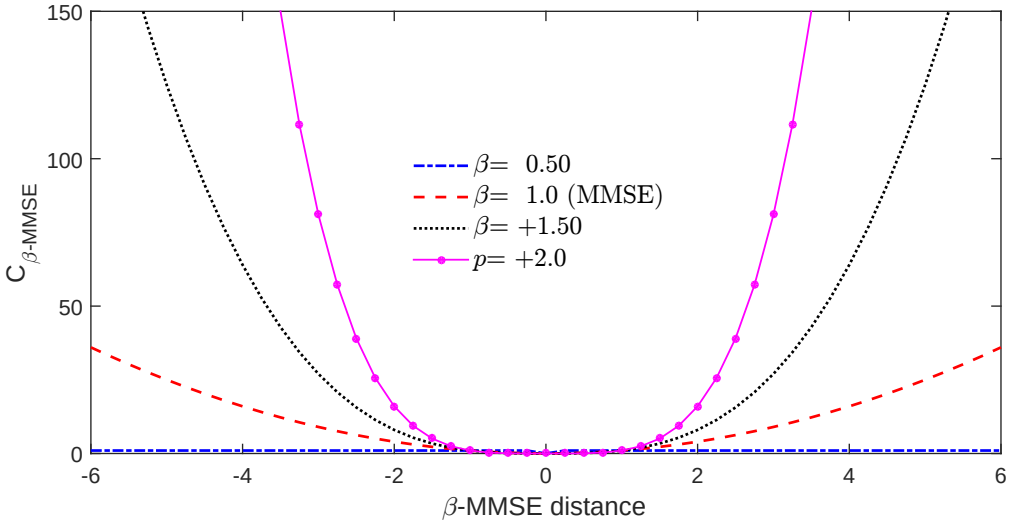
where exponent β is the parameter controlling the trade-off between speech distortion and noise reduction of the associated gain function. Inserting Eq. (4.26) to (4.15), gives

$$\mathbb{E}[\hat{N}|Z] = \left[\frac{\int_0^{\infty} n^{\beta}\, e^{-\frac{n^2}{\lambda^2}}\, I_0\left(\frac{2nz}{\lambda_{mx}^2}\right) dn}{\int_0^{\infty} e^{-\frac{n^2}{\lambda^2}}\, I_0\left(\frac{2nz}{\lambda_{mx}^2}\right) dn} \right]^{\frac{1}{\beta}}, \tag{4.27}$$

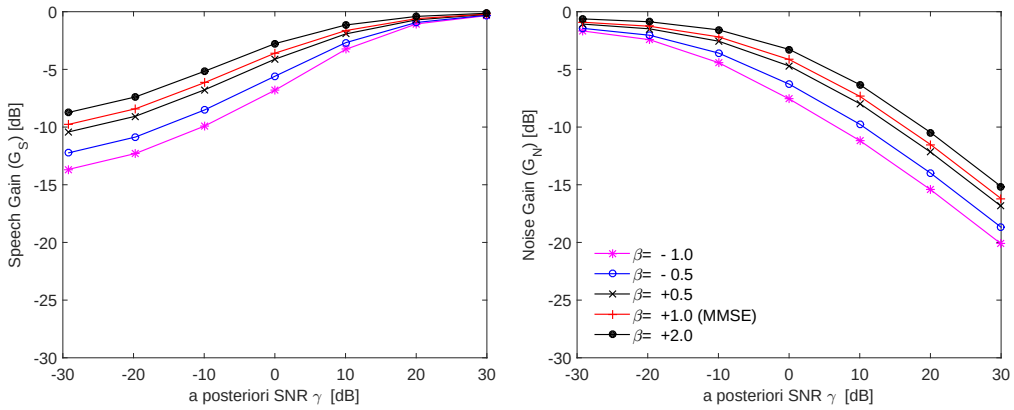
for which the β-MMSE noise gain corresponds to

$$G_N = \sqrt{\frac{\xi}{\gamma(\xi+1)}} \left[\Gamma\Big(\frac{\beta}{2}+1\Big)\Phi\Big(-\frac{\beta}{2},1;-\nu_n\Big)\right]^{\frac{1}{\beta}}. \tag{4.28}$$

Note that Eq. (4.19) is equivalent to the MMSE noise estimator (4.21) for the case when β=1.



(a) The plot of the β-MMSE distortion measure for varying exponent β.



(b) Speech gain (G_S) (c) Noise gain (G_N) Eq. (4.28)

Figure 4.2: The response of the generalized MMSE gain functions w.r.t. to the varying *a priori* and *a posteriori* SNRs for several values of β (4.33).

4.3.2.1 The case $\beta < 0$

For the negative values of the exponent β (< 0), we can write β= -$|\beta|$ and therefore the β-MMSE cost function (4.26) can be expressed by

$$C(n, \hat{n}) = \left(\frac{1}{n^{|\beta|}} - \frac{1}{\hat{n}^{|\beta|}} \right)^2 . \tag{4.29a}$$

$$= \frac{\left(\hat{n}^{|\beta|} - n^{|\beta|} \right)^2}{n^{2|\beta|}\ \hat{n}^{2|\beta|}} . \tag{4.29b}$$

Note that the numerator in Eq. (4.29b) represents the original β-MMSE cost function given in Eq. (4.26), whilst the denominator can be thought of as an approximation of the true noise power spectrum ($n^{2|\beta|}$). Due to this spectral power weighting, the estimator becomes more aggressive and gives larger estimation error for small noise spectral coefficients, i.e., spectral valleys.

Figure 4.2b and 4.2c plot the β-MMSE gain function responses for different values of β by using both the speech and noise estimators for varying the *a priori* and *a posteriori* SNRs, respectively. As observed from Fig. 4.2b, the speech estimator for a large value of β results low attenuation, while, the estimator reacts more aggressively if the exponent β reduces. In a similar way, the noise estimator for lower valued β provides smaller gain.

Importantly, for time-varying *a priori* and *a posteriori* SNRs, the slope in terms of *rate of rise* (ROR) for speech gain, and the *rate of fall* (ROF) for noise gain provides better understanding about the response of both the speech and noise estimators. Clearly, for increasing the a posterior SNR γ, the noise gain responds quickly whilst relatively slower variation is noticed in speech gain. It indicates that the noise estimator adapts the spectral changes more quickly than the speech estimator that inevitably reduces the probability of speech distortion. However, squaring the error might not be subjectively meaningful as MMSE estimator consider both positive and negative error a positive error even when actual estimation error is negative, and therefore, both the positive and negative errors need to be weighted accordingly.

The behavior of the β-MMSE estimator for $\beta < 0$ is similar to the Weighted Euclidean (WE) estimator. In the next subsection, we derive the noise estimator

by using the WE cost function in the next subsection.

4.3.3 Weighted Euclidean (WE) Measures

The perceptual effect of positive error i.e., the estimated magnitude is smaller than the true magnitude may differ with the negative error in the speech enhancement system and therefore both errors need not to be weighted equally. To overcome the above problems and shortcomings associated with the positive and negative errors in the MMSE estimator, several perceptually motivated distortion measures that give a more reasonable argument in support of human auditory system have been studied in [21]. Notably, the weighted Euclidean (WE), and the weighted COSH (WCOSH) distortion measures provide the weighting function that successfully hold the control over the estimation errors. This subsection derives the noise estimator by using the weighted Euclidean (WE) distortion measure while WCOSH in the next subsection.

The WE cost function is achieved by introducing the weighting function to the MMSE cost function in Eq. (4.14) as

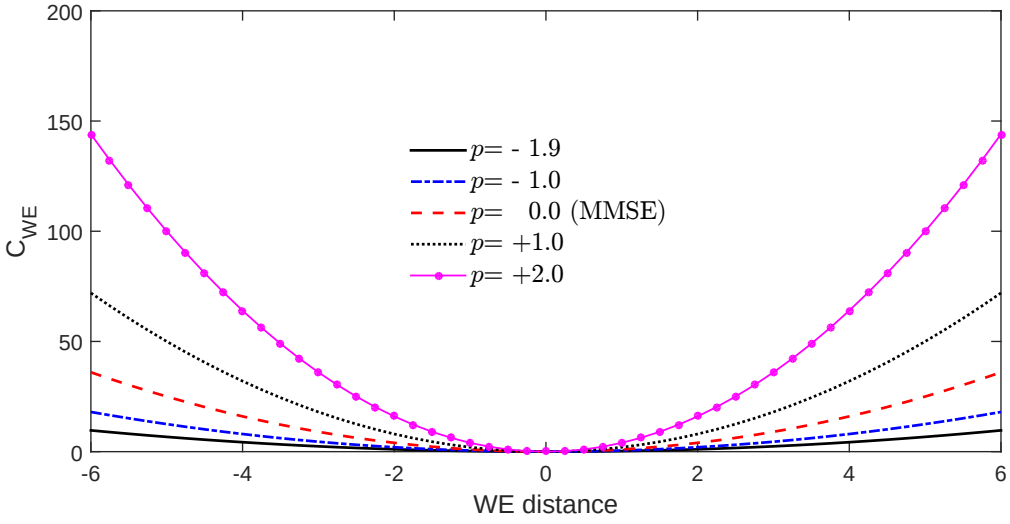
$$C(n,\hat{n}) = \frac{(n-\hat{n})^2}{n}. \tag{4.30}$$

The reason for using this weighted function is that it provides large estimation cost-error for smaller speech spectral coefficients (valley), whilst the smaller cost-error for large speech spectral coefficients (peaks) [21]. However, this weighting of the cost function given (4.30) may be sufficient for estimating the speech spectrum, it can be more important for estimating the noise spectral coefficients. The reason is that most of the state of the art noise estimation techniques fail to track the small noise spectral coefficients (noise spectral valley), and over-estimation occurs. As weighting function focuses more on the spectral valley, the speech distortion due to the over-estimation of the noise will be reduced. For estimating the noise, here the generalized weighting function suggested in [21], has been given by

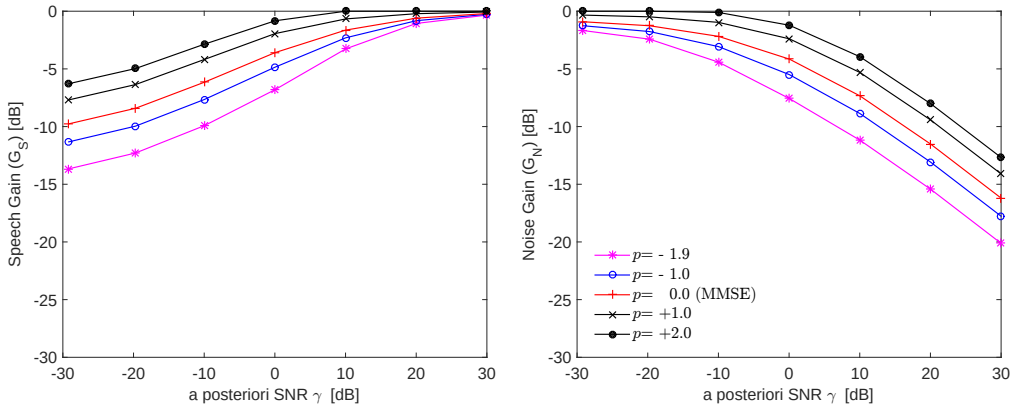
$$C(n,\hat{n}) = n^p(n-\hat{n})^2, \tag{4.31}$$

where, p is the exponent providing the appropriate weight to the estimator. As

$p \geq 0$ the cost function emphasizes noise spectral peaks whilst for $p \leq 0$, it focuses more on noise spectral valleys. The behavior of the WE estimator towards varying exponent p has been shown in Fig. 4.3.



(a) The plot of the WE distortion measure for varying exponent p.



(b) Speech gain (G_S) [21] (c) Noise gain (G_N)

Figure 4.3: The response of the (a) WE distortion measure, (b) speech gain, and (c) the noise gain functions w.r.t. to the varying *a priori* and *a posteriori* SNRs for several values of the weight exponent p.

By using Eq. (4.7), the WE noise estimator for the given cost function in Eq. (4.31) can be written, as

$$\mathbb{E}[\hat{N}|Z] = \frac{\int_0^\infty n^{p+1}\ f_{N|Z}(n|z)\ dn}{\int_0^\infty n^{p}\ f_{N|Z}(n|z)\ dn}, \tag{4.32}$$

By using the Gaussian statistical model, the expression can be evaluated in closed-

form and the corresponding WE estimator yields the noise gain by

$$G_N = \sqrt{\frac{\xi}{\gamma(\xi+1)}} \left[\frac{\Gamma\left(\frac{p+1}{2}+1\right) \Phi(-\frac{p+1}{2}, 1; -\nu_n)}{\Gamma\left(\frac{p}{2}+1\right) \Phi(-\frac{p}{2}, 1; -\nu_n)} \right] \quad \forall \ p > -2. \tag{4.33}$$

Note that, by setting $p = 0$, the traditional MMSE noise gain (4.21) is achieved as the confluent hyper-geometric function $\Phi(0, 1; -\text{x})$=1. Also, by letting $p = -1$, gives the cost function used in (4.30).

The WE based speech gain function [21, Eq. 18] responses for several values of the weight exponent p are shown in Fig. 4.3b and the noise gain function (4.33) responses in 4.3c, respectively. As we can see that, the response of the WE gain functions is almost similar to that of the β-MMSE noise gain functions given in Fig. 4.2. Additionally, the amount of attenuation provided by both methods seems to be dependent on the value of the power exponents β and p.

To take advantage of the perceptually motivated Bayesian estimators, next subsection derives the estimator by combining the weighting factors of both the β-MMSE estimator (4.26) and the WE estimator (4.31).

4.3.4 Weighted β-Order MMSE Measures

As both the β-MMSE (Eq. 4.26) and WE (Eq. 4.31) based estimators achieves considerable attention in estimating the speech DFT coefficients [20,21,127], [128] combining both the cost functions that includes all the possible variants of the MMSE estimators. The general form of the MMSE cost function is then written by combining these cost functions, as

$$C(n, \hat{n}) = n^p (n^\beta - \hat{n}^\beta)^2. \tag{4.34}$$

By using these two exponent parameters p and β, the generalized MMSE estimator will be derived which incorporates both the positive and negative errors. Importantly, the log-spectral amplitude (LSA) estimator which is correlated with the human auditory system can be achieved when $\beta \to 0$ [128]. For β with p=0, the β-MMSE cost function (4.26), whilst letting β=1 $\forall$ p, attains the WE cost function (4.31). By tuning these parameters appropriately, the estimator take

advantage of the perceptual or auditory interpretation.

Note that, estimating the noise DFT coefficients in modulation domain, this generalized form of the cost function will provide much flexibility of tuning both the parameters p and β, as the noise signal is slow varying compared to the speech in the DFT domain.

By using the above generalized cost function given in Eq. (4.34), the noise estimator can be achieved by minimizing the Bayesian risk function (4.7) by using the Eq. (4.9) and (4.10), as

$$\mathbb{E}[\hat{N}|Z] = \left[\frac{\int_0^\infty n^{\beta+p}\, e^{-\frac{n^2}{\lambda^2}}\, I_0\left(\frac{2nz}{\lambda_{mx}^2}\right) dn}{\int_0^\infty n^{p}\, e^{-\frac{n^2}{\lambda^2}}\, I_0\left(\frac{2nz}{\lambda_{mx}^2}\right) dn} \right]^{\frac{1}{\beta}}, \tag{4.35}$$

and by using [115, Eq. 6.631.1, 8.406.3, 9.212.1], the gain of the above noise estimator (4.35), evaluates to

$$G_N = \sqrt{\frac{\xi}{\gamma(\xi+1)}} \left[\frac{\Gamma\left(\frac{\beta+p}{2}+1\right)\, \Phi(-\frac{\beta+p}{2}, 1; -\nu_n)}{\Gamma\left(\frac{p}{2}+1\right)\, \Phi(-\frac{p}{2}, 1; -\nu_n)} \right]^{\frac{1}{\beta}}. \tag{4.36}$$

Since, $\Gamma(\mathsf{x})$ is valid only for its positive argument x, the restriction of selecting the parameters β and p has been imposed, i.e.,

$$\beta + p > -2, \quad \text{and} \quad p > -2. \tag{4.37}$$

Considering these limitations, the above noise gain function includes all possible combinations of β and p to achieve the MMSE and WE estimators as shown in Table 4.1.

Fig. 4.4 represents the gain response of the β-order MMSE noise estimator variants by using different combinations of the β and p. It is observed that the estimator gives a similar response as that of the general MMSE (Fig. 4.2c) and GWE noise gain (Fig. 4.3c) functions. The advantage of using this β-order MMSE noise estimators that, it provides a larger dynamic range by having different combinations of the exponents. In this noise gain functions plotted in Fig. 4.4 clearly show that, the increase in the exponent β, results the gain

Table 4.1: The cost functions with their respective noise gains (G_N) for several existing MMSE estimator variants.

β-order MMSE	$(p,\ \beta)$	Noise gain (G_N)
$\left[n^p(n^\beta - \hat{n}^\beta)^2\right]$	p=0, β=1 (MMSE)	$\Gamma(\frac{3}{2})\ \sqrt{\frac{\xi}{\gamma(\xi+1)}}\ \Phi(-\frac{1}{2},1;-\nu_n)$,
	p=0, β (GMMSE)	$\sqrt{\frac{\xi}{\gamma(\xi+1)}}\left[\Gamma\left(\frac{\beta}{2}+1\right)\Phi(-\frac{\beta}{2},1;-\nu_n)\right]^{\frac{1}{\beta}}\ \forall\ \beta > -2,$
	p=-1, β=1 (WE)	$\sqrt{\frac{\xi}{\gamma(\xi+1)}}\ \frac{1}{\Gamma(\frac{1}{2})}\ \frac{e^{(\frac{\nu_n}{2})}}{I_0(\frac{\nu_n}{2})}$, [125, Eq. A1.31b]
	p, β=1 (GWE)	$\sqrt{\frac{\xi}{\gamma(\xi+1)}}\left[\frac{\Gamma\left(\frac{p+1}{2}+1\right)\ \Phi(-\frac{p+1}{2},1;-\nu_n)}{\Gamma\left(\frac{p}{2}+1\right)\ \Phi(-\frac{p}{2},1;-\nu_n)}\right]\ \forall\ p > -2,$

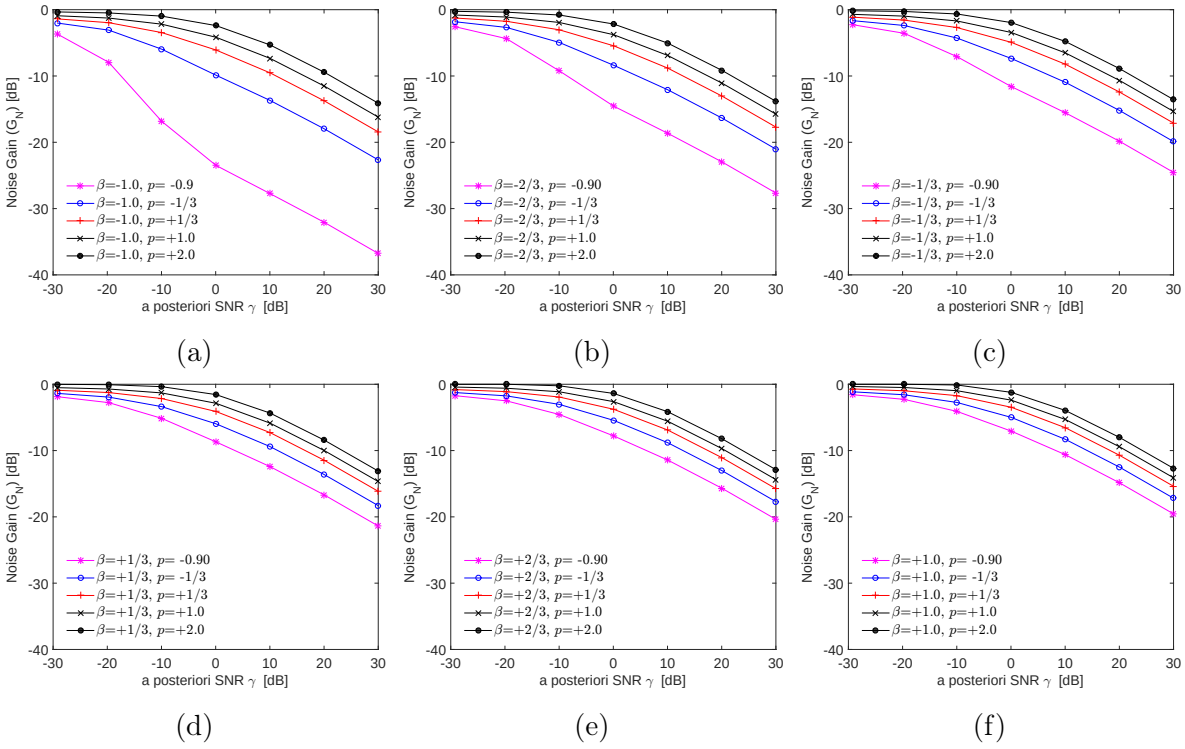


Figure 4.4: The response of the β-order MMSE gain functions using (a) $\beta = -1.0$, (b) $\beta = -\frac{2}{3}$, (c) $\beta = -\frac{1}{3}$, (d) $\beta = +\frac{1}{3}$, (e) $\beta = +\frac{2}{3}$, and (f) $\beta = +1.0$ for varying p.

increment whilst reducing the exponent p, decreases.

4.3.5 Weighted COSH (WCOSH) Measures

An asymmetric distortion measure Itakura-Saito (IS) [129] has been successfully implemented in speech recognition system. Motivated by that, [130] combines two forms of the IS distortion measure that results in a new but symmetrical distortion measure, as

$$C(n,\hat{n}) = \frac{1}{2}\left(\frac{n}{\hat{n}} + \frac{\hat{n}}{n}\right) - 1. \tag{4.38}$$

Note that this distortion measure is well-known by COSH distortion measure and a variant of the well-known IS distortion measure. As explained in [21], the COSH measure penalizes large estimation errors more heavily but penalizes small estimation error that is nearly identical to the log spectral distortion suggested in [22]. Therefore, the generalization of this COSH measure is provided in [21], by representing the cost function, as

$$C(n,\hat{n}) = \left[\left(\frac{n}{\hat{n}} + \frac{\hat{n}}{n}\right) - 1\right] n^{p}, \tag{4.39}$$

where p is the weighting parameter and not necessarily limited to be an integer. As shown in Fig. 4.5a, the $WCOSH$ distortion measure provides different variant of the $COSH$ measure by varying the parameter p.

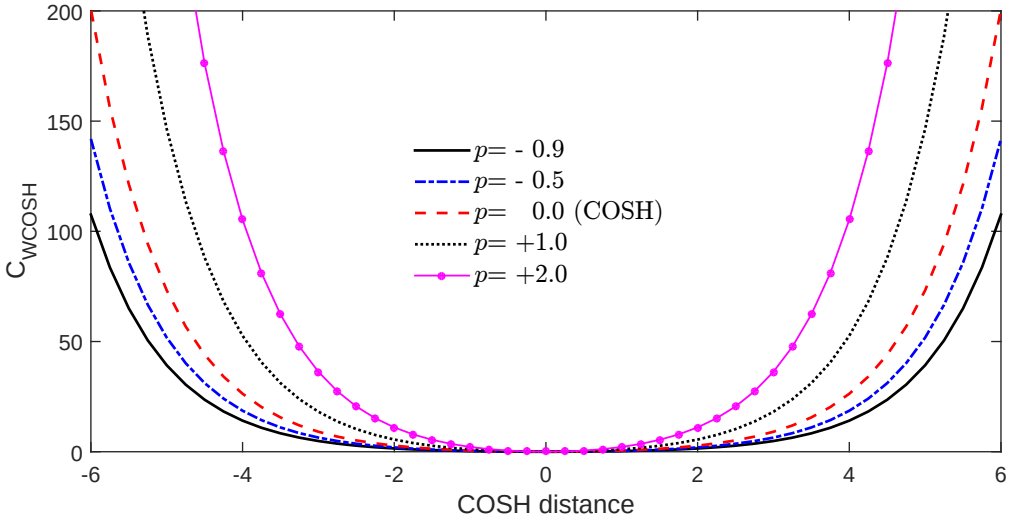
Minimization of the the risk function with the $WCOSH$ distortion measure (4.39), gives

$$\mathbb{E}[\hat{N}|Z] = \frac{\displaystyle\int_0^{\infty} n^{p+1}\, f_{N|Z}(n|z)\, dn}{\displaystyle\int_0^{\infty} n^{p-1}\, f_{N|Z}(n|z)\, dn}, \tag{4.40}$$

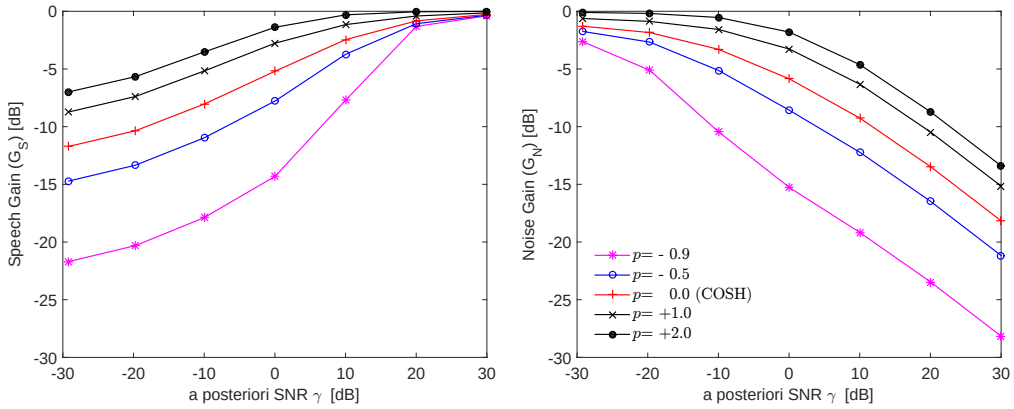
and the associated WCOSH noise estimator is obtained by solving the Eq. (4.40), i.e.,

$$\mathbb{E}[\hat{N}|Z] = \sqrt{\frac{\xi}{\gamma(\xi+1)} \frac{\Gamma\left(\frac{p+3}{2}\right)}{\Gamma\left(\frac{p+1}{2}\right)} \frac{\Phi(-\frac{p+1}{2},1;-\nu_n)}{\Phi(-\frac{p-1}{2},1;-\nu_n)}}\, Z \quad \forall \;\; p > -1. \tag{4.41}$$

Fig. 4.5 plots the response of the noise gain. As observed, for $p<0$, the noise gain decreases whilst smaller difference in the gain response is noticed for $p>0$.



(a) The plot of the WE distortion measure for varying exponent p.



(b) Speech gain (G_S) [21, Eq. 34] (c) Noise gain (G_N)

Figure 4.5: The response of the (a) WCOSH distortion measure with its derived speech gain and noise gain functions w.r.t. to the varying *a priori* and *a posteriori* SNRs for several values of the weighting exponent p (4.41).

4.4 Summary

This chapter presented the perceptually motivated Bayesian estimators for tracking the all time-varying noise signals in the modulation domain. This is motivated by the fact that the Bayesian estimators seemly represent the estimation error by incorporating the perceptual aspects of the speech signals in the frequency domain. However, investigative experimental findings from chapter 3 clearly indicate that the modulation domain has a capability to improve the speech intelligibility by providing a more predictive non-stationary spectral variations for both the speech and noise signals compared to the frequency domain.

Therefore, exploring the properties of the modulation domain towards track-

ing the noise DFT coefficients more closely, this chapter derived the family of noise estimators by utilizing the perceptually motivated Bayesian theory. More specifically, the squared-error based MMSE estimator treated both the positive and negative squared errors equally whilst the weighted Euclidean (WE) method entertains the spectral valleys more that the spectral peaks. To encompass all, the weighted β-order MMSE noise estimator is analytically derived which represents the family of the MMSE estimator by incorporating all perceptual aspects for the speech and noise signals in both the frequency and modulation DFT domains.

As noticed, the exponent β used in the squared error cost function (4.26) allows the appropriate tuning in correlating the human auditory system. For example, $\beta \to 0$ represents the well known Log-Spectral Amplitude (LSA) estimation which is successfully established in correlating the human auditory system. Moreover, Eq. (4.31) gives the standard MMSE estimator by using $p{=}0$ whilst the negative value ($p{<}0$) represents the WE estimator and the estimator penalize the spectral valley more than the spectral peaks.

To investigate the effectiveness of these perceptually motivated Bayesian noise estimators in the modulation domain, next chapter will have various comprehensive experiments.

$\sim * \sim$

Chapter 5

Modulation Domain Bayesian Results and Analysis

An Equation means nothing, unless it expresses a thought of God.

–Sir Srinivasa Ramanujan.

5.1 Introduction

This chapter provides their performance and behavior of the Bayesian motivated noise methods in the modulation domain. The approximations of these noise estimators are analytically derived here to validate the usability of the Bayesian estimator's for the single-channel noise estimation. To differentiate the performance of MMSE and COSH based noise methods, the estimators are divided into two categories. In the first category, the MMSE based noise estimators are considered while the second category considers the performance of COSH based noise methods.

Details of the experimental setup is given in chapter 3, i.e., the stimuli of 12 phonetically balanced sentences consisting of six different male and six different female speakers from the TIMIT database [44]. Four different noise sources with varying stationarity namely, stationary white noise, long term stationary factory noise, non-stationary babble noise and highly non-stationary street noise are added to the speech at a wide range of input SNRs.

5.2 Weighted β-MMSE Noise Estimator

5.2.1 β-MMSE with Limiting Case ($\beta \to 0$)

It is argued that the β-order MMSE speech gain mathematically represent the log-spectral amplitude (LSA) estimator when β approaches to zero ($\beta \to 0$) [127]. Similarly for the noise estimator derived by using the β-MMSE, we provide the mathematical validation as the noise gain itself manifests as LSA noise estimator when $\beta \to 0$.

Expressing the noise gain derived from Eq. (4.28) in the form of

$$G_N = \sqrt{\frac{\xi}{\gamma(\xi+1)}}\ \exp\left\{\frac{1}{\beta}\ln\Big[\Gamma\Big(\frac{\beta}{2}+1\Big)\Big] + \frac{1}{\beta}\ln\Big[\Phi\Big(-\frac{\beta}{2},1;-\nu_n\Big)\Big]\right\}, \tag{5.1}$$

where, ξ, γ are the *a priori* and *a posteriori* SNRs respectively while β is the exponent used in the β-MMSE cost function. By using [115, Eq. 8.342.1], (5.1) can be represented by

$$G_N = \sqrt{\frac{\xi}{\gamma(\xi+1)}}\ \exp\left\{-\frac{\Upsilon}{2}+\frac{1}{\beta}\sum_{a=2}^{\infty}(-1)^a\,\frac{1}{a}\Big(\frac{\beta}{2}\Big)^a\,\zeta(a)+\frac{1}{\beta}\ln\Big[\Phi\Big(-\frac{\beta}{2},1;-\nu_n\Big)\Big]\right\}, \tag{5.2}$$

where, Υ is the Euler's constant, $\zeta(.)$ is the Weierstrass's zeta function [115],

$$\text{and}\quad \ln\Big[\Gamma(\mathsf{x}+1)\Big] = -\Upsilon\mathsf{x}+\sum_{a=2}^{\infty}(-1)^a\,\frac{\mathsf{x}^a}{a}\,\zeta(a)\ \ \forall\ \ |\mathsf{x}|<1, \tag{5.3}$$

where a is an integer. After applying the limitation to β ($\beta \to 0$), the noise gain translated to

$$\lim_{\beta\to 0}\ G_N = \sqrt{\frac{\xi}{\gamma(\xi+1)}}e^{-\gamma_e/2}\,\exp\left[\lim_{\beta\to 0}\frac{\ln\Big[\Phi\Big(-\frac{\beta}{2},1;-\nu_n\Big)\Big]}{\beta}\right], \tag{5.4}$$

Interestingly, the term $\left[\lim_{\beta\to 0},\frac{\ln\left[\Phi\left(-\frac{\beta}{2},1;-\nu_n\right)\right]}{\beta}\right]$ appears as the indeterminate form of $0/0$, and therefore, differentiation of both the numerator and denominator from

the L'Hospital's rule gives

$$\lim_{\beta\to 0} \; G_N = \sqrt{\frac{\xi}{\gamma(\xi+1)}} e^{-\gamma_e/2} \exp\left[\lim_{\beta\to 0} \frac{\frac{\delta}{\delta\beta}\left[\Phi\Big(-\frac{\beta}{2},1;-\nu_n\Big)\right]}{\Phi\Big(-\frac{\beta}{2},1;-\nu_n\Big)}\right] \tag{5.5}$$

The numerator of Eq. (5.5) represents the series expansion of the confluent hypergeometric function [20], and can be written as

$$\lim_{\beta\to 0} \frac{\delta}{\delta\beta}\left\{\Phi\Big(-\frac{\beta}{2},1;-\nu_n\Big)\right\} = -\frac{1}{2}\sum_{b=1}^{\infty}\frac{1}{b}\frac{\big(-\nu_n\big)^b}{b!}, \tag{5.6}$$

whilst,

$$\lim_{\beta\to 0}\left\{\Phi\Big(-\frac{\beta}{2},1;-\nu_n\Big)\right\} = 1, \tag{5.7}$$

where b is an integer. Letting Eq. (5.6), (5.7) in to Eq. (5.5), yields

$$\lim_{\beta\to 0} \; G_N = \sqrt{\frac{\xi}{\gamma(\xi+1)}} \exp\left[-\frac{\gamma_e}{2} - \frac{1}{2}\sum_{b=1}^{\infty}\frac{1}{b}\frac{\big(-\nu_n\big)^b}{b!}\right]. \tag{5.8}$$

Additionally, from [115, Eq. 8.211.1, 8.214.1], we get

$$-\gamma_e - \ln(\mathsf{x}) - \sum_{b=1}^{\infty}\frac{1}{b}\frac{\big(-\mathsf{x}\big)^b}{b!} = \int_{\mathsf{x}}^{\infty}\frac{e^t}{t}dt \quad \forall \;\; \mathsf{x}>0 \tag{5.9}$$

which transforms Eq. (5.8) in to

$$\lim_{\beta\to 0} \; G_N = \sqrt{\frac{\xi}{\gamma(\xi+1)}} \exp\left[\frac{1}{2}\ln(\nu_n) + \frac{1}{2}\int_{\nu_n}^{\infty}\frac{e^{-t}}{t}dt\right] \tag{5.10a}$$

$$= \frac{1}{\xi+1}\exp\left[\frac{1}{2}\int_{\nu_n}^{\infty}\frac{e^{-t}}{t}dt\right] \tag{5.10b}$$

which is the LSA noise gain function.

From Fig. 5.1, the β-MMSE noise gain for limiting $\beta \to 0$ holds the log-spectral amplitudes (LSA) estimator's characteristics and gives the similar performances. Clearly, for decreasing the β the gain values are decreasing, whilst noise estimator is more aggressive when β increases.

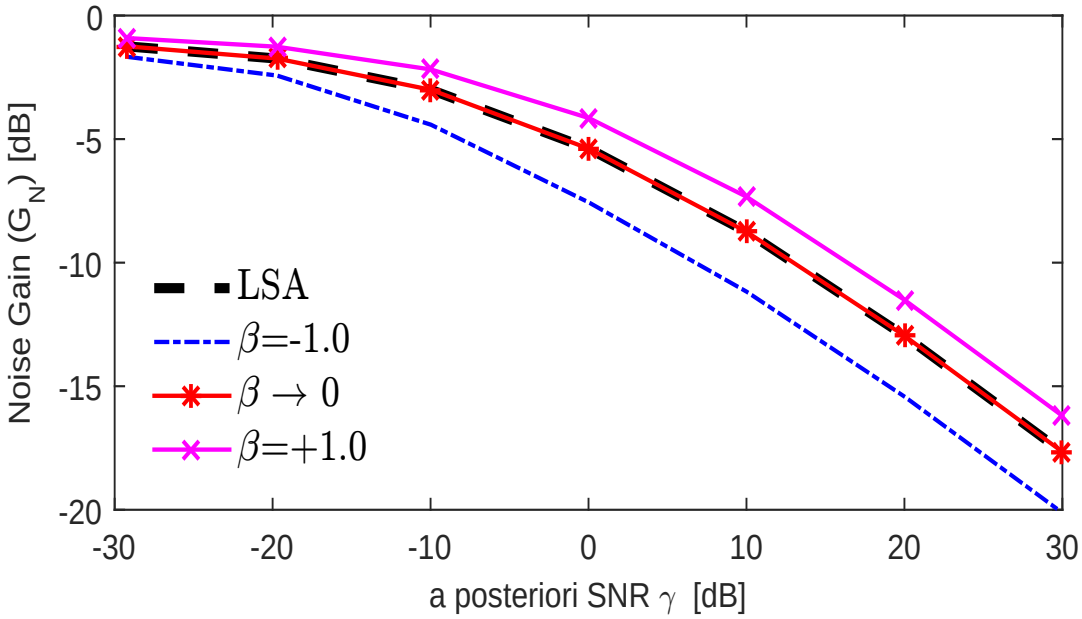


Figure 5.1: The LSA noise gain (5.10b) plots with the weighted β-MMSE noise gain function for $\beta \to 0$, by providing β=0.001, in Eq. (5.1).

5.2.2 Weighted β-MMSE with Large *a priori* SNR ($\gamma \gg 1$)

It is well known that ν_n is the function of γ and ξ, as $\nu_n = \frac{\gamma}{\xi(\xi+1)}$, that means when $\gamma \to \infty$, ν_n also approaches to ∞. In this asymptotic relation of ν_n with γ, the confluent hyper-geometric function is approximated [125, Eq. A1.16b], by

$$\lim_{\nu_n \to \infty} \Phi(\mathrm{a}, 1; -\nu_n) \approx \frac{\nu_n^{-\mathrm{a}}}{\Gamma(1-\mathrm{a})}. \tag{5.11}$$

For the noise gain approximation, the gain function derived using weighted β-order MMSE cost function is repeated here for convenience

$$G_N = \sqrt{\frac{\xi}{\gamma(\xi+1)}} \left[\frac{\Gamma\left(\frac{\beta+p}{2}+1\right) \Phi(-\frac{\beta+p}{2}, 1; -\nu_n)}{\Gamma\left(\frac{p}{2}+1\right) \Phi(-\frac{p}{2}, 1; -\nu_n)} \right]^{\frac{1}{\beta}}. \tag{5.12}$$

Substituting the approximation of confluent hyper-geometric function for large value of $\gamma(\gg 1)$, the noise gain function given in Eq. (5.12) represents

$$\lim_{\nu_n \to \infty} G_N \approx \sqrt{\frac{\xi}{\gamma(\xi+1)}} \left[\frac{\Gamma\left(\frac{\beta+p}{2}+1\right) \Gamma\left(\frac{p}{2}+1\right) (\nu_n)^{\left(\frac{\beta+p}{2}\right)}}{\Gamma\left(\frac{p}{2}+1\right) \Gamma\left(\frac{\beta+p}{2}+1\right) (\nu_n)^{\left(\frac{p}{2}\right)}} \right]^{\frac{1}{\beta}}. \tag{5.13}$$

After simplification, Eq. (5.13) yields

$$\lim_{\nu_n \to \infty} G_N \approx \sqrt{\frac{\xi}{\gamma(\xi+1)}} \left(\frac{{\nu_n}^{(\frac{\beta+p}{2})}}{{\nu_n}^{(\frac{p}{2})}} \right)^{\frac{1}{\beta}}, \tag{5.14}$$

and, by letting $\nu_n = \frac{\gamma}{\xi(\xi+1)}$, it gives the well known Wiener noise gain by

$$\lim_{\nu_n \to \infty} G_N \approx \frac{1}{\xi+1}, \tag{5.15}$$

which is the well known Wiener noise gain function.

5.2.3 Modulation FFT Size Considerations

As described earlier, the modulation framework selection for improving the speech quality and intelligibility is motivated by the fact that, different modulation based applications use different FFT size. As smaller FFT size provides higher intelligibility [27], whilst Speech based applications such as hearing aid devices prefer smaller FFT size. Applications such as Speech coding, on the other hand, prefer better quality over intelligibility [46] and, the selection of modulation framework may differ by having a larger FFT size. In a similar way, the performance of the noise estimator in the modulation domain may differ from estimator to estimator.

Therefore, to understand the effect of the modulation FFT size with the exponents β and p values used in the weighted β-order MMSE noise estimator (4.36) for estimating the noise spectral amplitudes, we have conducted investigative experiments with various modulation FFT lengths [1] (16 to 256) and varying β and p values.

In these experiments, 12 phonetically balanced sentences of six different male and six different female speakers from the TIMIT corpus [44], and four types of noise signals covering all practical scenario of time-varying stationarity from NOISEX-92 [45] have been included at 0, 5, 10, 15, and 20dB input SNR levels. The intelligibility based performance of the weighted β-MMSE noise estimator is analyzed by using PESQ score, and the results are shown in Fig. 5.2 for stationary white noise, Fig. 5.3 for long term stationary factory noise, Fig. 5.4

[1] The acoustic FFT size 512, acoustic frame shift 6.25% and modulation frame shift 50% are considered from subsection 3.5.3 to avoid the results complexity.

for heavy street noise, and Fig. 5.5 for non-stationary babble noise. Besides, the mean STOI score based performance of the weighted β-MMSE noise estimator is presented in the APPENDIX A.1 section.

It is clear from Fig. 5.2 that, the modulation FFT size plays an important role towards achieving better speech intelligibility. For example, increasing the FFT size, i.e., 256 and above, the overall performance of the weighted β-MMSE noise estimator degraded as intelligibility is downgraded. Whilst is lowering the FFT size (<16), the similar performance is noticed. Although for highly noisy conditions (input SNR $\leq$ 8dB), the FFT range to give better intelligibility is 32 to 128 where estimator performs satisfactorily. This is especially when input SNR

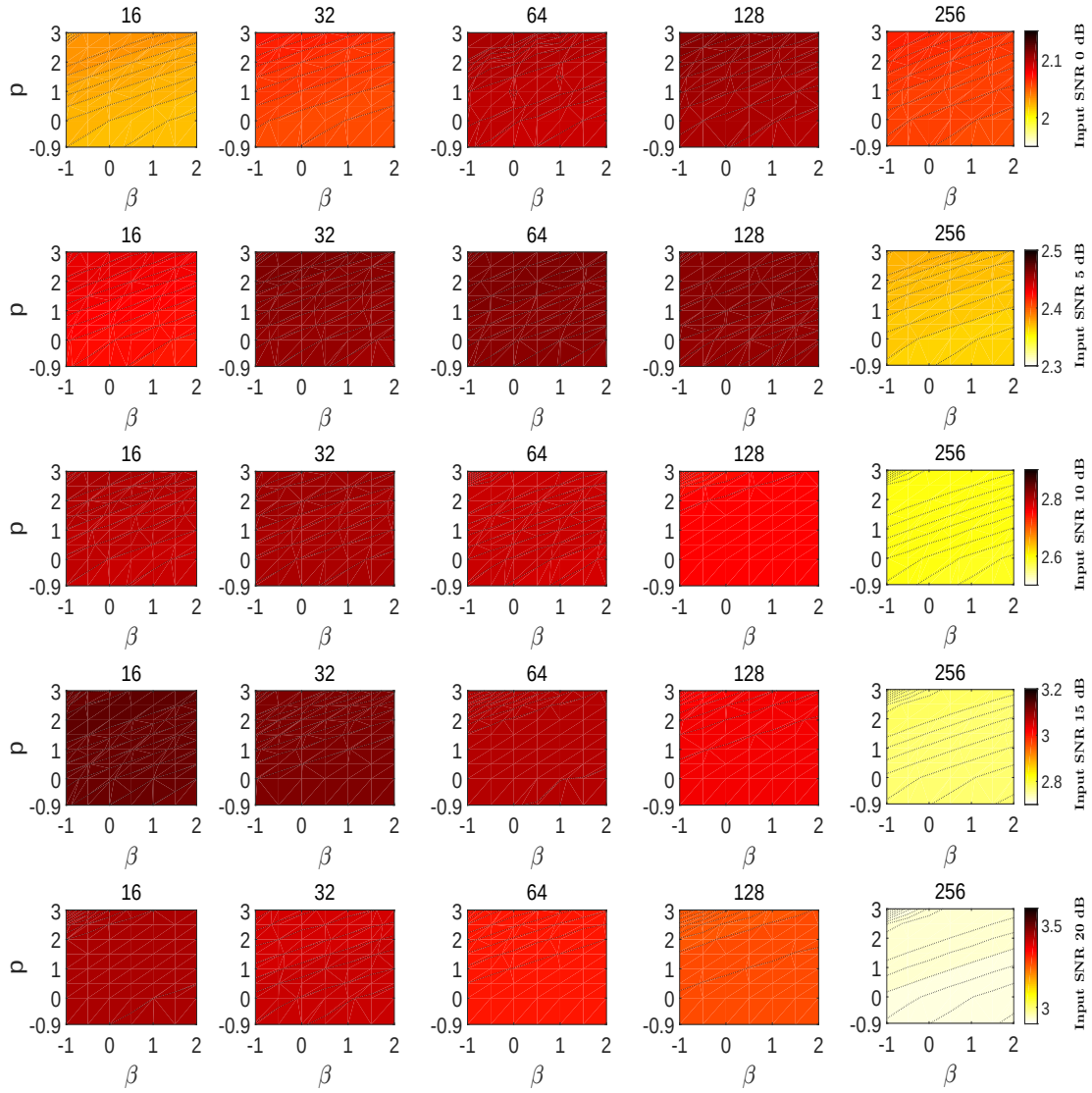


Figure 5.2: The stationary white noise based mean intelligibility (PESQ) score for varying modulation FFT size (MFS 50%), β, and p values.

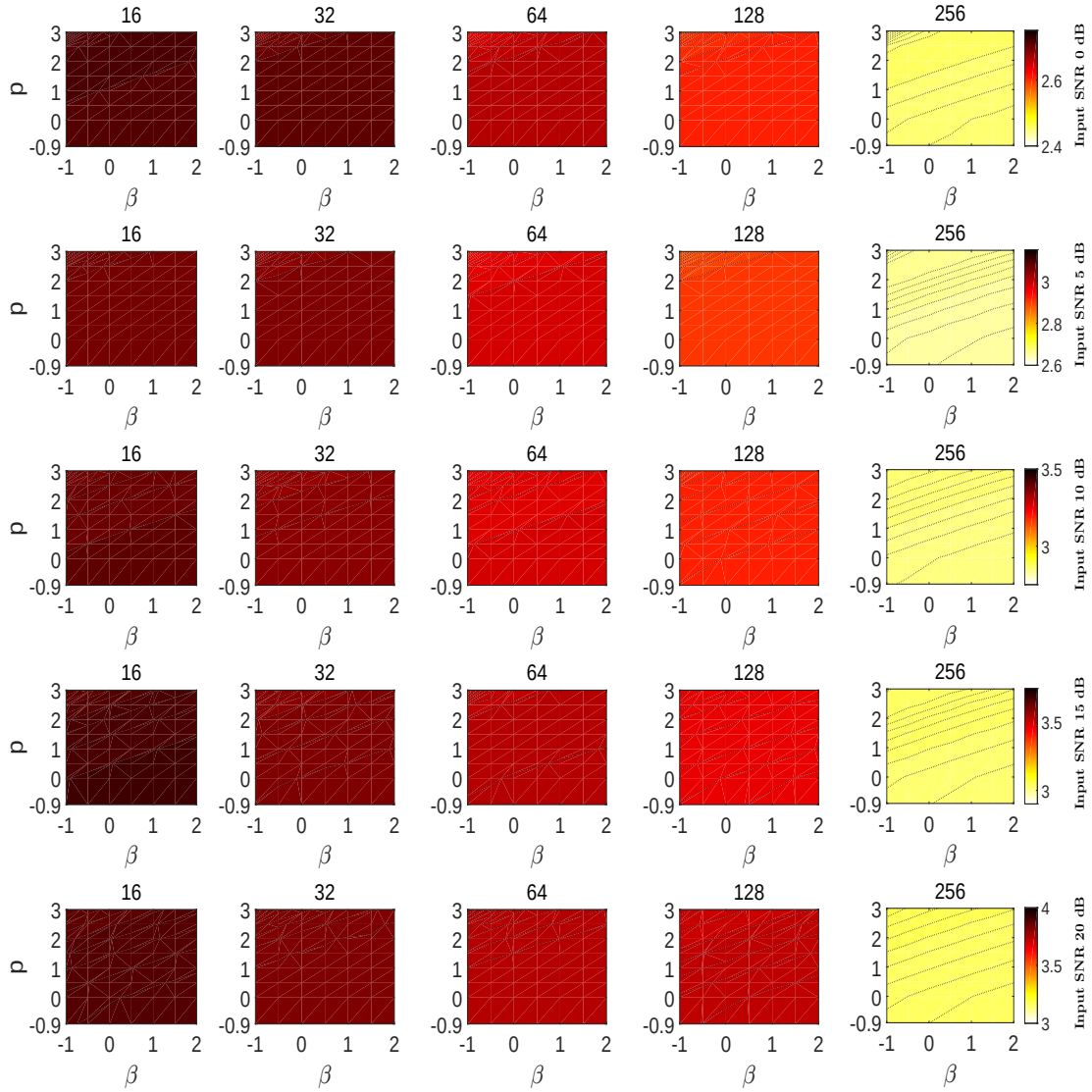


Figure 5.3: The long term stationary factory noise based mean intelligibility (PESQ) score for varying modulation FFT size (MFS 50%), β, and p values.

is low, whilst increasing the input SNR $\geq$ 8dB), the similar performance can be noticed by lowering the FFT size.

Contrary to this, the estimator's performance for long-term stationary factory noise is different as shown in Fig. 5.3. For this, the FFT size of 16 and 32 promise better results for all the SNR level. As we can see, the FFT size 16 allows estimator to track the noise more appropriately, whilst MFFT size 32 gives similar results as we get by using the MFFT size of 16.

Interestingly, the estimator's performances for highly non-stationary noise signals like heavy street noise shown in Fig. 5.4, or babble noise plotted in Fig. 5.5, are similar as both nose types are difficult to estimate because of their highly

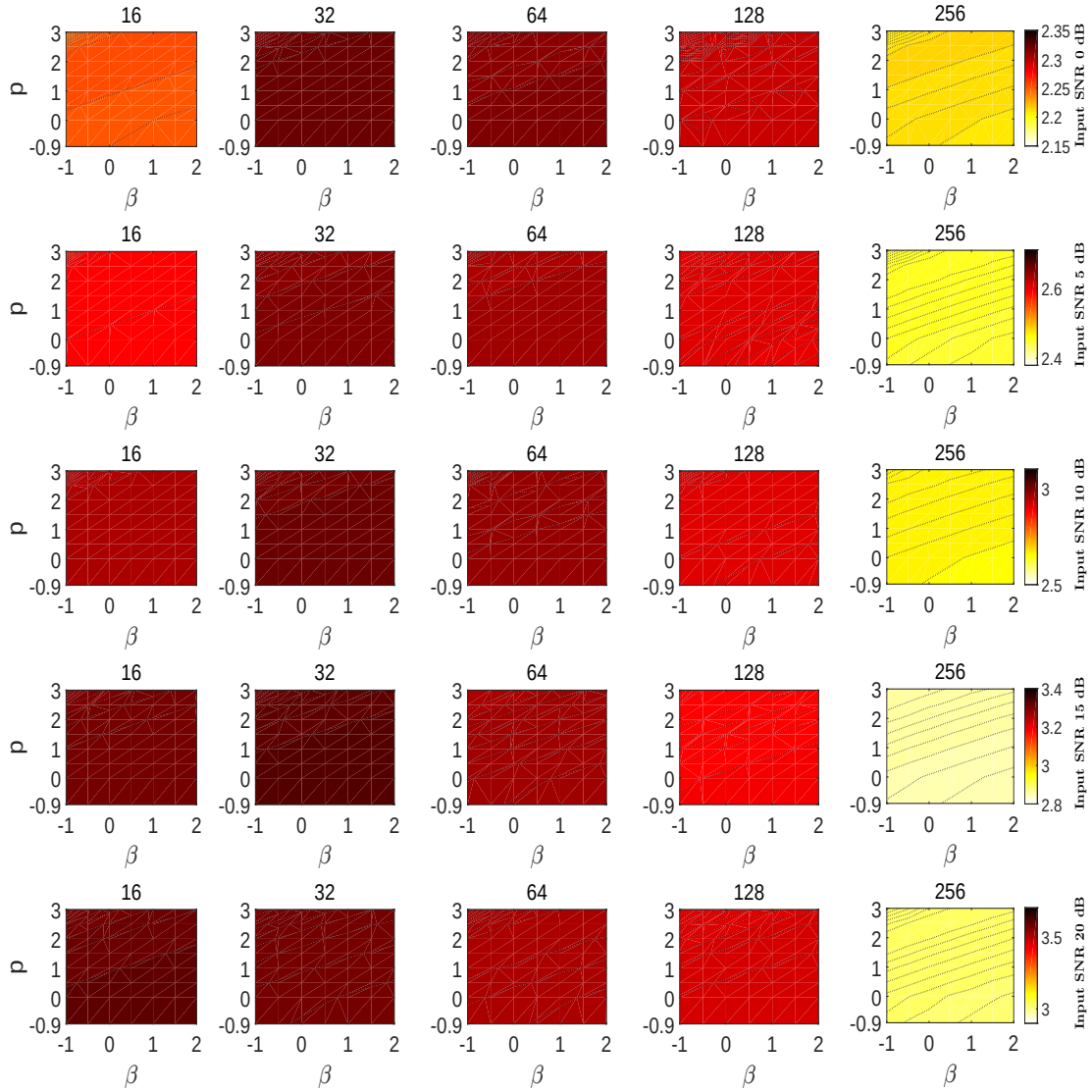


Figure 5.4: The heavy street noise based mean intelligibility (PESQ) score for varying modulation FFT size (MFS 50%), β, and p values.

non-stationary nature. More specifically, the intelligibility of the enhanced speech degraded by street noise can be achieved by using the MFFT size of 32 or 64 for the input SNR level $\leq$ 10dB while increasing the input SNR (*geq* 10dB) estimator gives better intelligibility by using the MFFT of 32 or 16.

In succeeding subsection 5.2.4, we describe the results achieved by using the weighted β-MMSE noise estimator, whilst subsections 5.2.5, and 5.3.3 describe the role of β and the exponent p used in the weighted β-MMSE noise estimator.

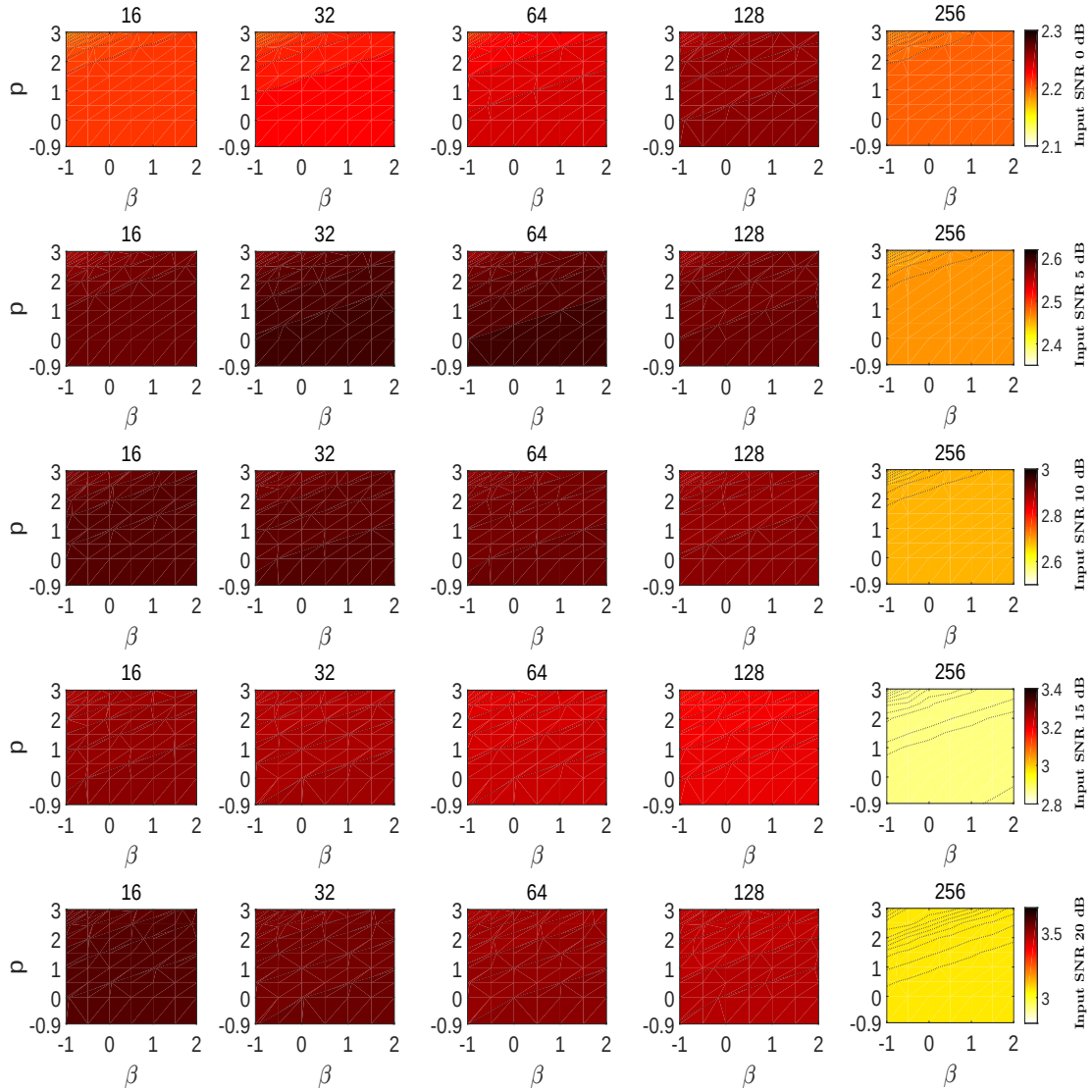


Figure 5.5: The non-stationary babble noise based mean intelligibility (PESQ) score for varying modulation FFT size (MFS 50%), β, and p values.

5.2.4 Wβ-MMSE Speech & Noise Estimators Performance

To illustrate the behavior of the Bayesian motivated weighted β-MMSE noise estimator in modulation domain, the noise estimator's performance has been compared with the speech estimator derived in [128, Eq. 7], which evaluates

$$G_N = \frac{\sqrt{\nu_k}}{\gamma} \left[\frac{\Gamma\left(\frac{\beta+p}{2}+1\right) \Phi(-\frac{\beta+p}{2}, 1; -\nu_k)}{\Gamma\left(\frac{p}{2}+1\right) \Phi(-\frac{p}{2}, 1; -\nu_k)} \right]^{\frac{1}{\beta}}. \tag{5.16}$$

For convenience, the noise estimator derived in the previous chapter (4.36) is repeated here as

$$G_N = \sqrt{\frac{\xi}{\gamma(\xi+1)}} \left[\frac{\Gamma\left(\frac{\beta+p}{2}+1\right)\ \Phi(-\frac{\beta+p}{2},1;-\nu_n)}{\Gamma\left(\frac{p}{2}+1\right)\ \ \Phi(-\frac{p}{2},1;-\nu_n)} \right]^{\frac{1}{\beta}}. \tag{5.17}$$

where,

$$\nu_k = \frac{\xi}{\xi+1}\gamma, \qquad \nu_n = \frac{\nu_k}{\xi^2} \ = \ \frac{\gamma}{\xi\big(\xi+1\big)}. \tag{5.18}$$

The methods that assess the overall quality of the enhanced speech here is divided by its intelligibility based measures i.e., PESQ and STOI scores, and segmental signal-to-noise (SNRseg) based measure.

Throughout the presentation of results in this chapter, the evaluation will be mainly focused on the aspects such as the nature of the noise estimator (including both β and p values) towards modulation domain. To facilitate a modulation based comparison between both speech and noise estimators, the acoustic FFT size and frame shift are fixed with 512 and 6.25%. The modulation FFT size 32, and frame shift 50% are motivated from the above subsections i.e., 3.5.3 and 5.2.3.

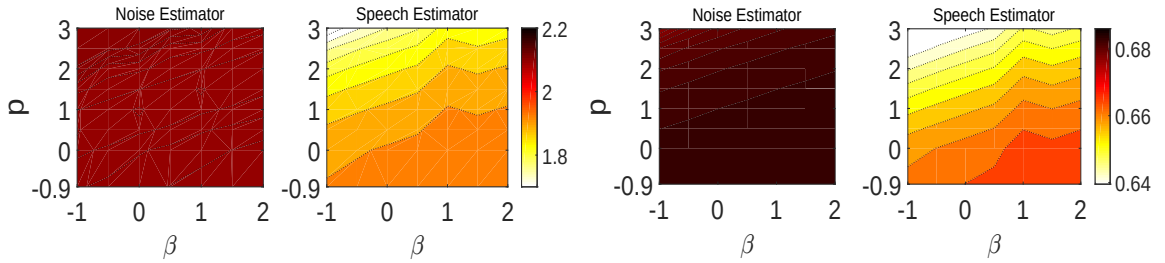
Intelligibility Measures

The stationary noise (white): The intelligibility based results for stationary white noise degraded speech are shown in Fig. 5.6, where the PESQ scores are given in sub-figs. 5.6a, 5.6c, 5.6e, 5.6g, 5.6i, and the STOI scores are in 5.6b, 5.6d, 5.6f, 5.6h, and 5.6j. The results clearly show that the noise estimator achieves considerable and consistent speech intelligibility improvement across the input SNR range as compared to the speech estimator. As we can see that the PESQ based performance of the speech is getting better by reducing the exponent p values whilst contrary to this, the speech intelligibility is improving by using the noise estimator. The factor β on the other hand, allows both the speech and noise estimators to work satisfactory for all values. Although, when β value is high, the range of estimator's applicability towards the exponent p is slightly increases as we get increased range of the exponent p values.

Besides, as the STOI scores are shown in sub-figs. 5.6b, 5.6d, 5.6f, 5.6h, and 5.6j, it is well noted that the noise estimator gives better intelligibility of the speech as compared to the speech estimator. Specifically, for lower input SNR conditions ($\leq$ 10dB). The similar performance can be noticed while changing the exponent p. For instance, by using the speech estimator, intelligibility increases for lowering the exponent p values. The different observation can be seen for input SNR 20dB where speech estimator performs better relatively. One reason may be that, for input SNR 20dB, mostly the speech components presents which make easier to estimate the speech components rather than the noise spectral amplitudes.

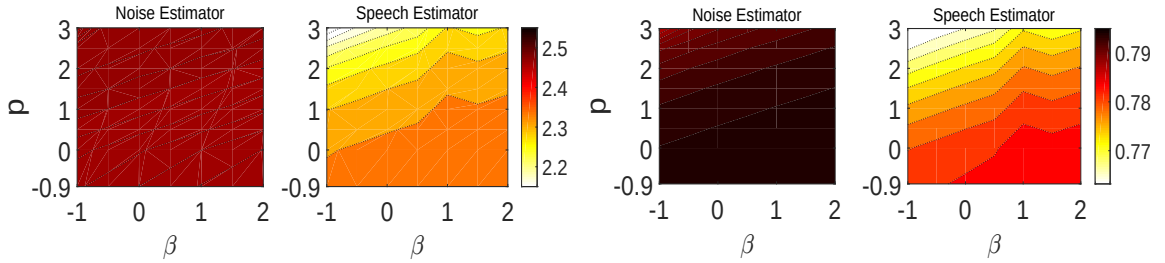
The long-term stationary noise (factory): Fig. 5.7 shows the results plot for the factory noise in which sub-figs. 5.7a, 5.7c, 5.7e, 5.7g, 5.7i represent the PESQ scores, whilst the STOI scores are given in sub-figs. 5.7b, 5.7d, 5.7f, 5.7h, 5.7j. The estimation process of the semi-stationary factory noise is similar up-to the input SNR 10dB as the noise estimator dominates by providing the better speech intelligibility but the effectiveness of the noise estimator decreases when input SNR increases ($\geq$ 10dB). It may be because, the factory noise contains the varying spectral amplitudes that influence the behavior of the speech spectral properties. This influence is enormous for lower input SNR and, as a consequence nature of the speech spectral coefficients affected largely. As shown in Fig. 5.7, the experimental results clearly indicate that the perfomance of both the speech and noise estimator differs from the stationary white noise. Moreover, the role of the β and exponent p values in noise estimator are similar to the stationary white noise. For instance, the lower value of p delivers better intelligibility for all values of β. Although, increasing the β values conveys larger range of exponent p values in noise estimation, the speech estimator allows better estimation by using the positive β values, especially when the input SNR is low.

The highly non-stationary noise (street & babble): The PESQ and STOI score based results for heavy street noise and babble noise are plotted in Fig. 5.8 and Fig. 5.9, respectively. The main issues are of interest is how fast the noise estimator reacts to the noise spectral variations in modulation domain,



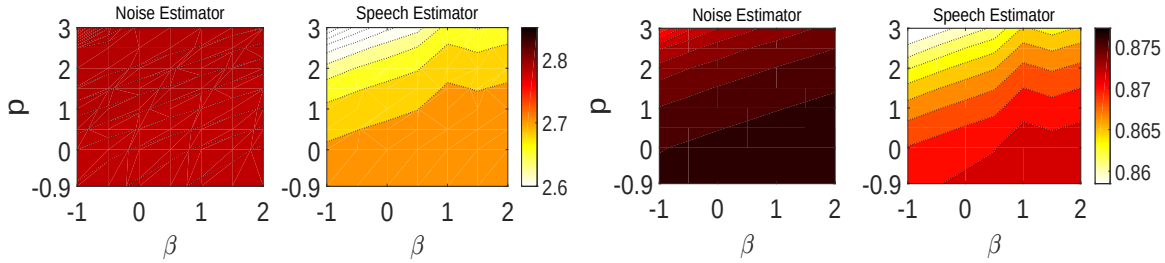
(a) PESQ (0 dB).

(b) STOI (0 dB).



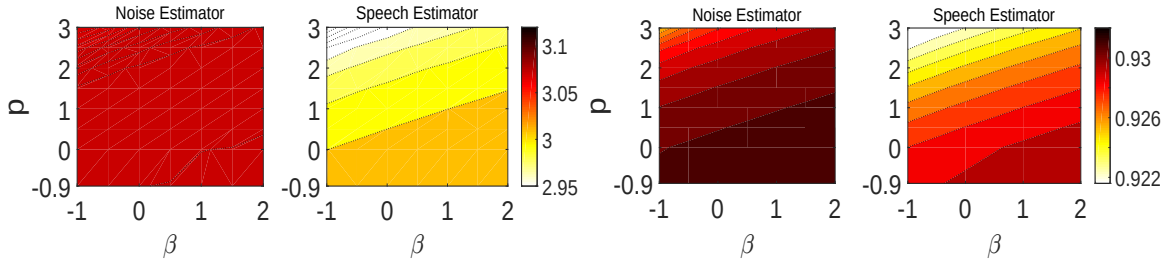
(c) PESQ (5 dB).

(d) STOI (5 dB).



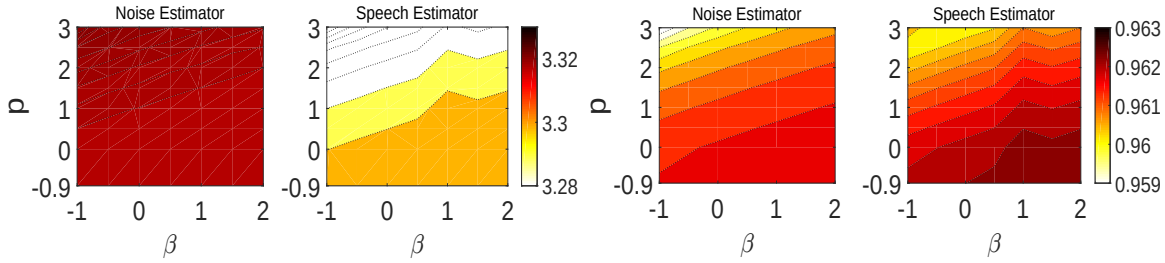
(e) PESQ (10 dB).

(f) STOI (10 dB).



(g) PESQ (15 dB).

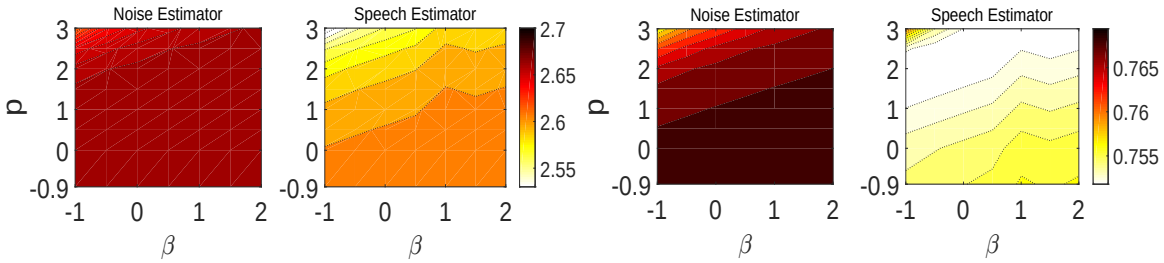
(h) STOI (15 dB).



(i) PESQ (20 dB).

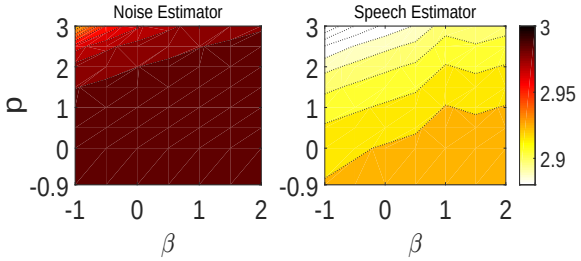
(j) STOI (20 dB).

Figure 5.6: The stationary white noise based mean PESQ (a, c, e, g, i) and STOI scores (b, d, f, h, j) of enhanced speech achieved by using both the speech and noise estimators.

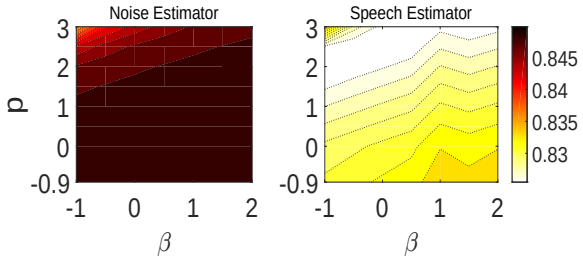


(a) PESQ (0 dB).

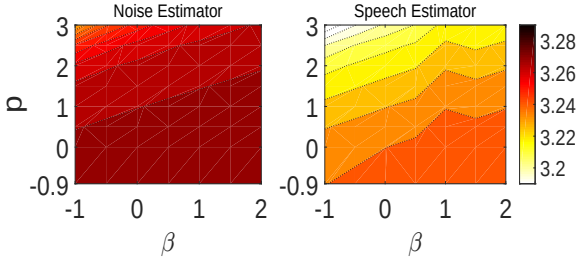
(b) STOI (0 dB).



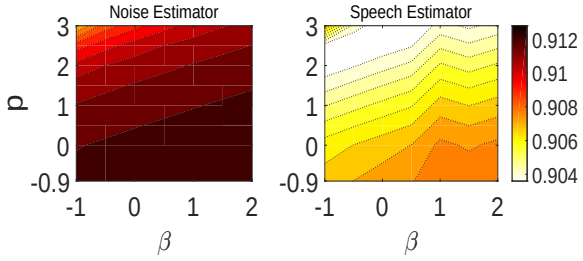
(c) PESQ (5 dB).



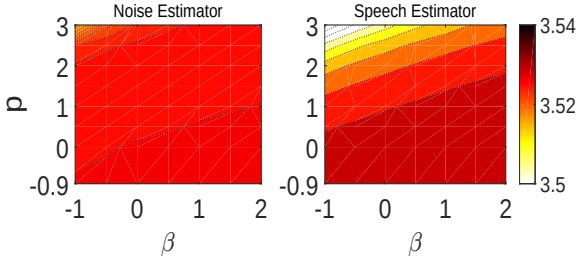
(d) STOI (5 dB).



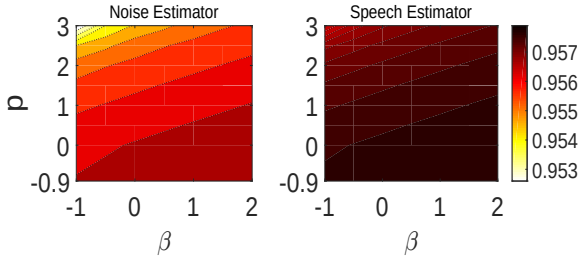
(e) PESQ (10 dB).



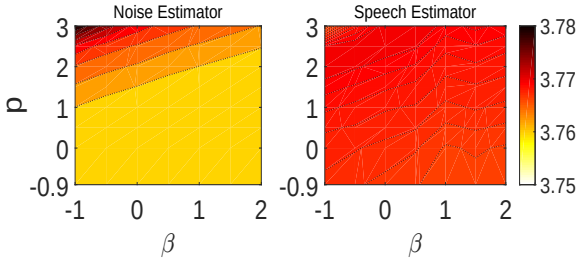
(f) STOI (10 dB).



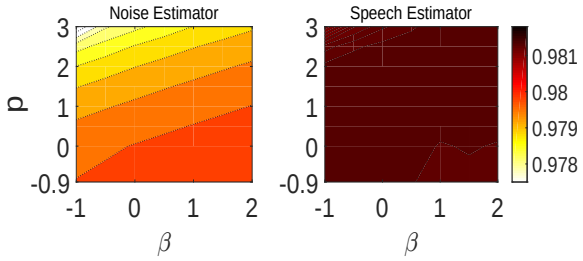
(g) PESQ (15 dB).



(h) STOI (15 dB).

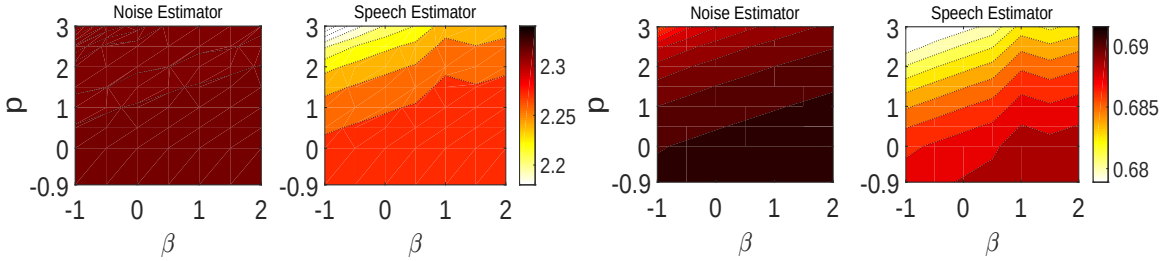


(i) PESQ (20 dB).



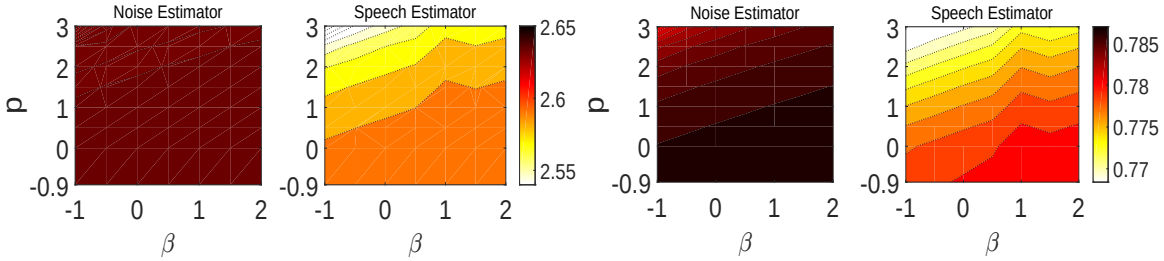
(j) STOI (20 dB).

Figure 5.7: The factory noise based mean PESQ (a, c, e, g, i) and STOI scores (b, d, f, h, j) of enhanced speech achieved by using both the speech and noise estimators.



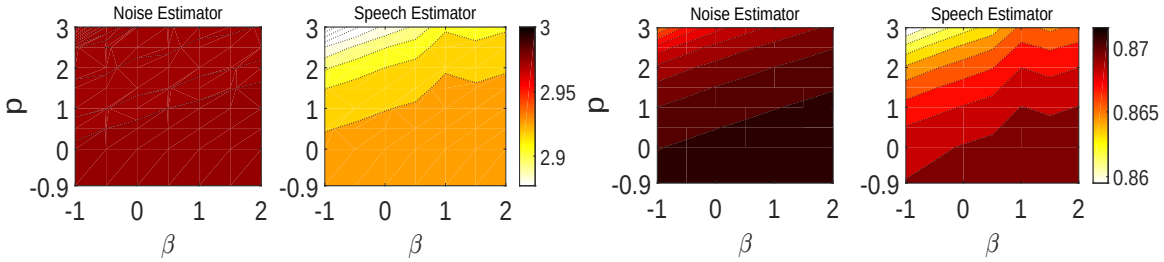
(a) PESQ (0 dB).

(b) STOI (0 dB).



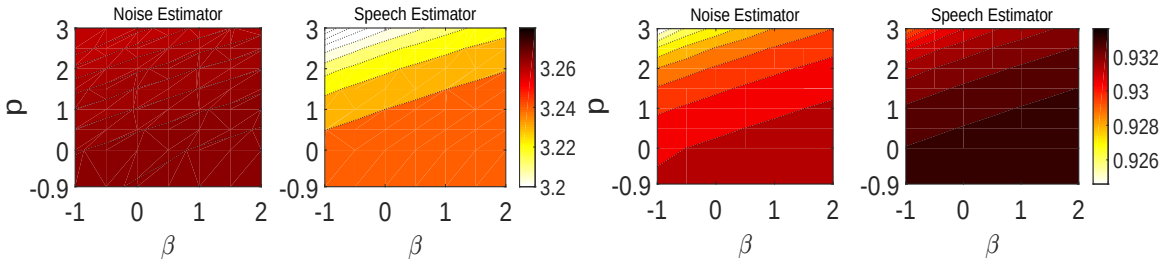
(c) PESQ (5 dB).

(d) STOI (5 dB).



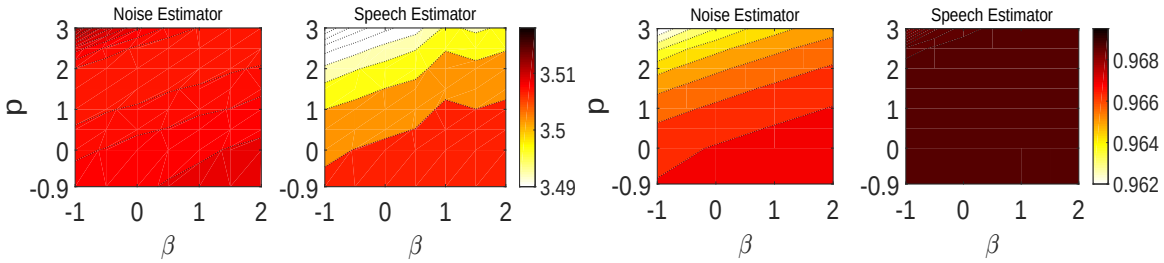
(e) PESQ (10 dB).

(f) STOI (10 dB).



(g) PESQ (15 dB).

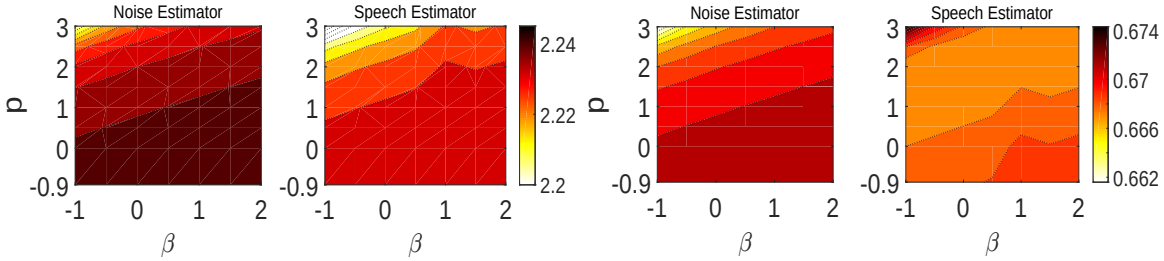
(h) STOI (15 dB).



(i) PESQ (20 dB).

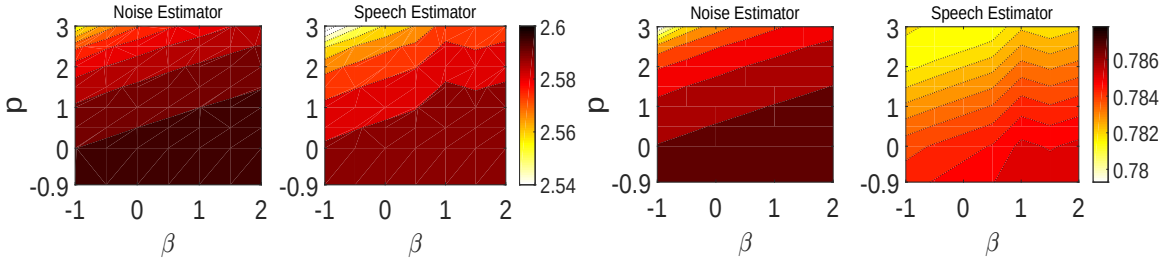
(j) STOI (20 dB).

Figure 5.8: The heavy street noise based mean PESQ (a, c, e, g, i) and STOI scores (b, d, f, h, j) of enhanced speech achieved by using both the speech and noise estimators.



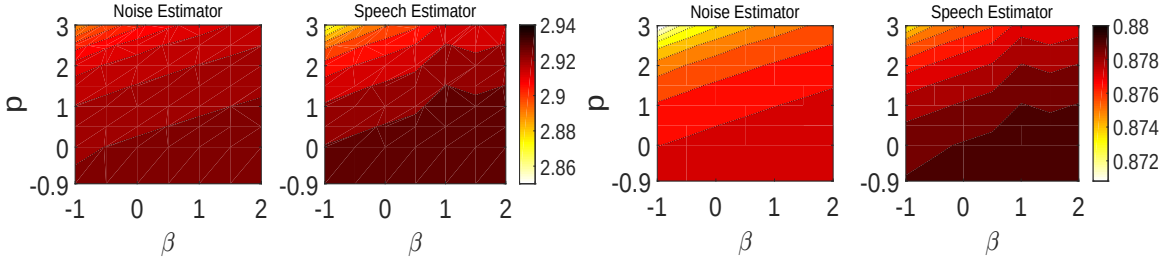
(a) PESQ (0 dB).

(b) STOI (0 dB).



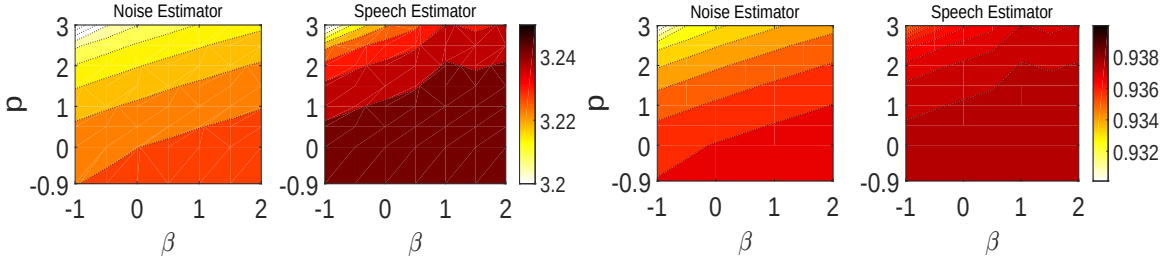
(c) PESQ (5 dB).

(d) STOI (5 dB).



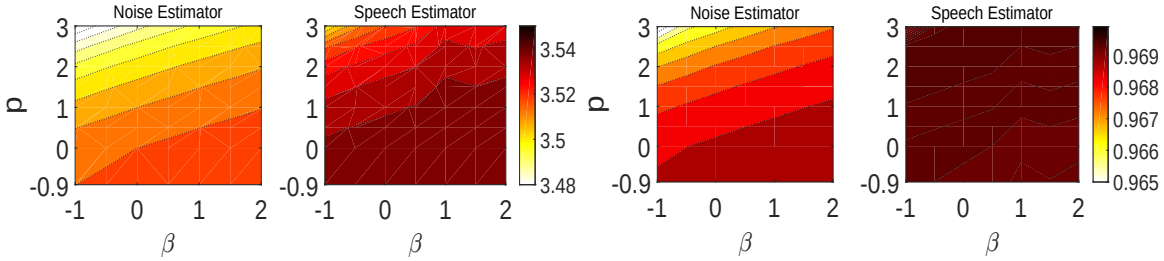
(e) PESQ (10 dB).

(f) STOI (10 dB).



(g) PESQ (15 dB).

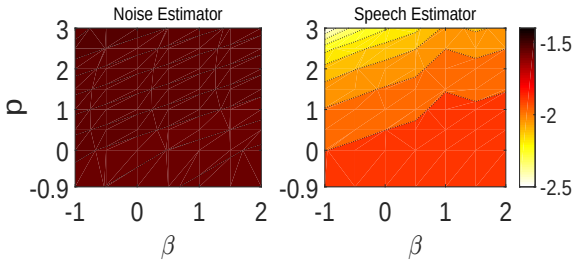
(h) STOI (15 dB).



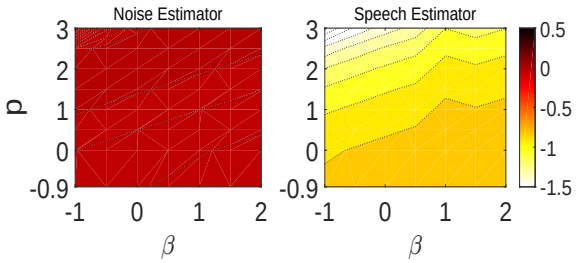
(i) PESQ (20 dB).

(j) STOI (20 dB).

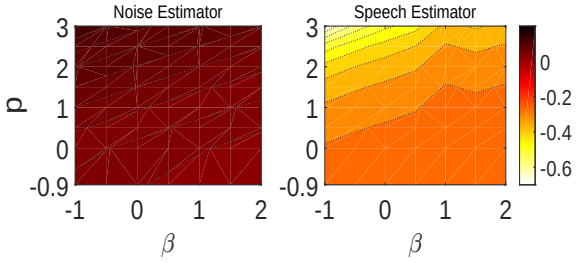
Figure 5.9: The non-stationary babble noise based mean PESQ (a, c, e, g, i) and STOI scores (b, d, f, h, j) of enhanced speech achieved by using both the speech and noise estimators.



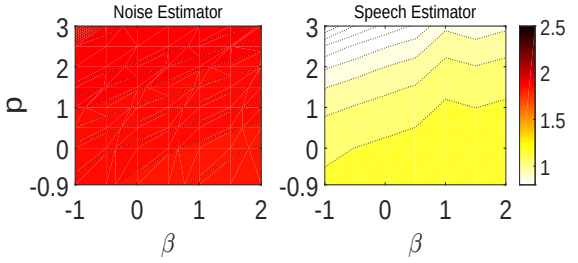
(a) white noise (0 dB).



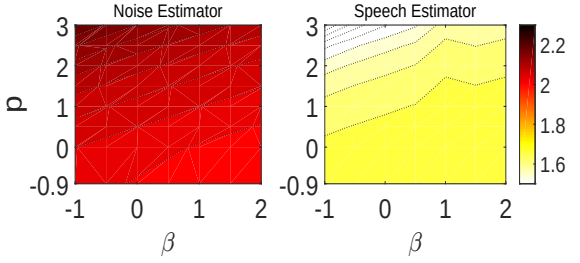
(b) factory noise (0 dB).



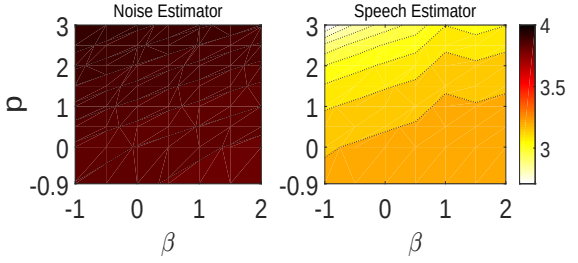
(c) white (5 dB).



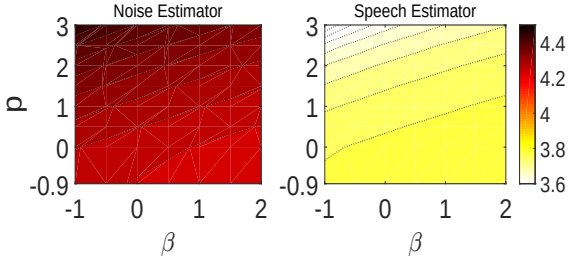
(d) factory (5 dB).



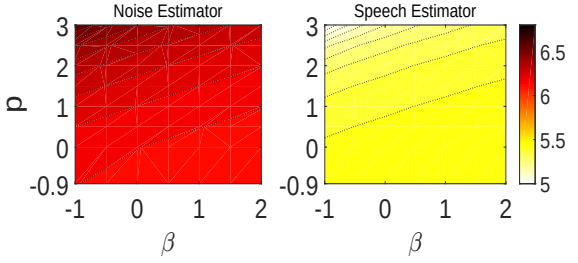
(e) white (10 dB).



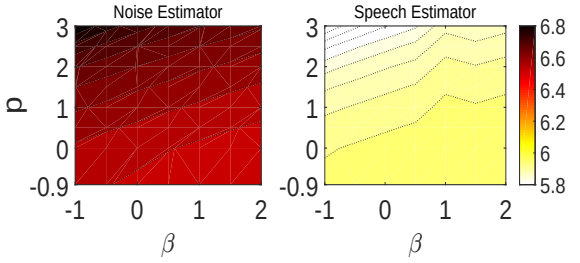
(f) factory (10 dB).



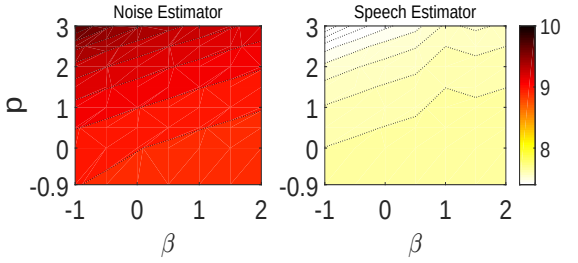
(g) white (15 dB).



(h) factory (15 dB).

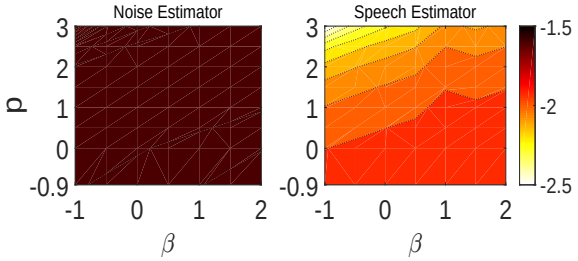


(i) white (20 dB).

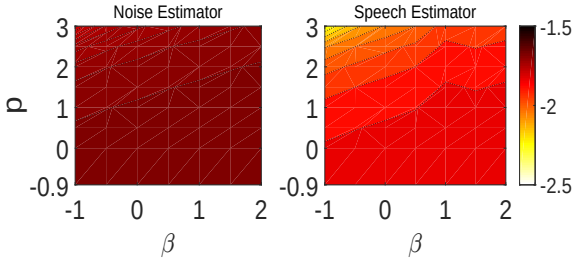


(j) factory (20 dB).

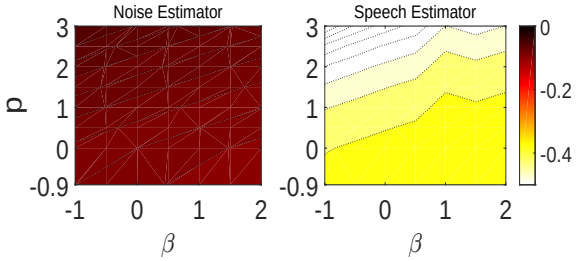
Figure 5.10: The stationary white (a, c, e, g, i) and factory (b, d, f, h, j) noise based mean segmental SNR of enhanced speech achieved by using both the speech and noise estimators.



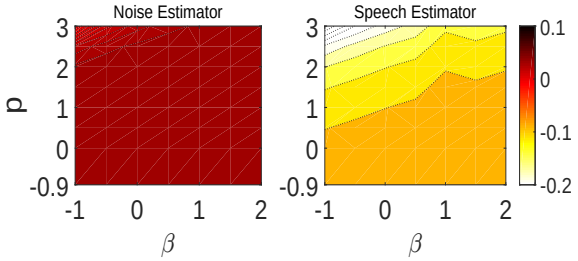
(a) street noise (0 dB).



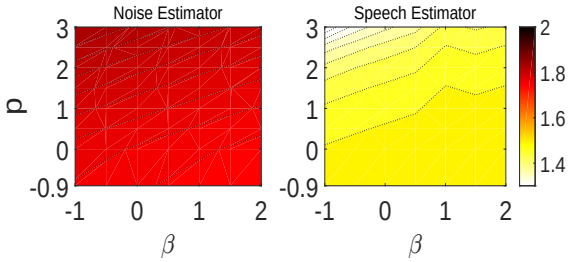
(b) babble noise (0 dB).



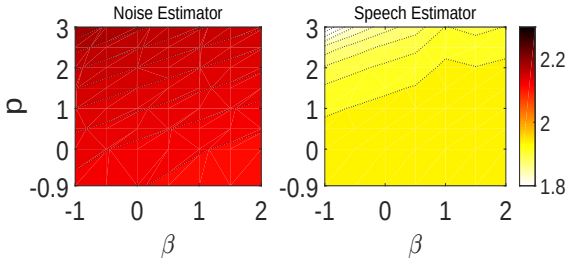
(c) street (5 dB).



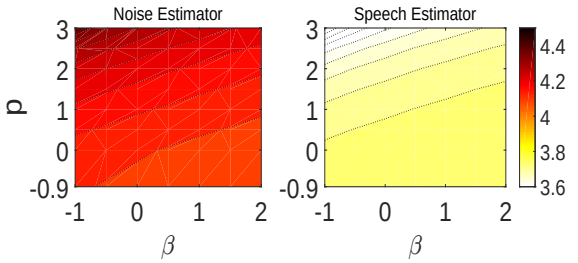
(d) babble (5 dB).



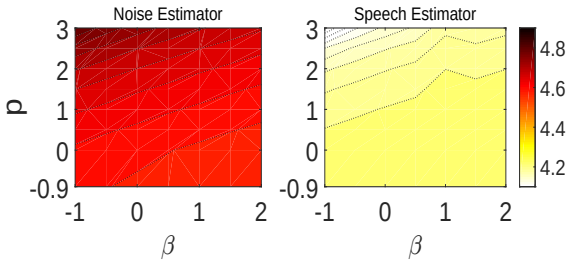
(e) street (10 dB).



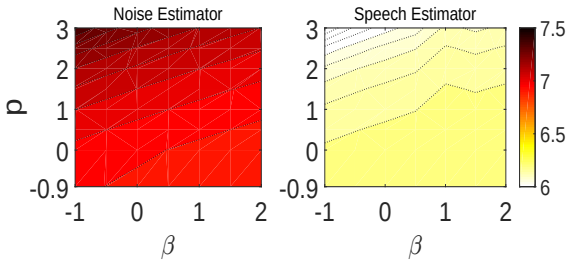
(f) babble (10 dB).



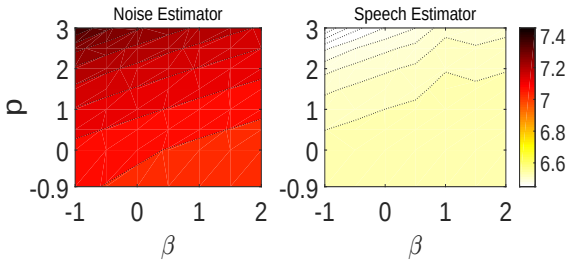
(g) street (15 dB).



(h) babble (15 dB).



(i) street (20 dB).



(j) babble (20 dB).

Figure 5.11: The heavy street (a, c, e, g, i) and non-stationary babble (b, d, f, h, j) noise based mean segmental SNR of enhanced speech achieved by using both the speech and noise estimators.

because more adaptive the noise estimator, more accurately the noise will be estimated. For example, the speech estimator does not provide much PESQ improvement compared to the noise estimator in a heavy noise case ($\leq$ 10dB). However for low input SNR, the speech estimator may assume the varying noise spectral components as speech which results in an underestimation (residual) of the noise and causes the loss of intelligibility.

As shown in Figs. 5.8, and 5.9, improving the speech intelligibility of the enhanced speech using non-stationary noise signals is a challenging task especially for low input SNR conditions. The reason may be that the babble noise consists of the multiple speech components from the neighboring speakers and street noise has a mixture of different noise signals. Due to these non-predictable noise spectral components, noise has various peaks distributed randomly in the spectral domain and the estimator assumes the noise as a speech, which gives a large estimation error especially for low input SNR conditions. On the other hand, tracking the speech spectral component becomes easier by lowering the noise level ($\geq$10dB). It is because, most of the spectral components are correlated with the speech signal and the by using the noise estimator, the probability of noise miss-detection will become high as noise spectral components become weaker. This effect can be easily seen from these figs.

Segmental SNR Measures

The segmental SNR (SNRseg) is an extension of the traditional (total) SNR and is designed to measure more accurately the quality of the enhanced speech. The motivation for this measure is to emphasize the effect of noise in the low energy speech segment, which is more sensitive to noise compared to the high energy ones. The results based on the segmental SNR (SNRseg) Measure of stationary white noise and factory noise are presented in Fig. 5.10, whilst Fig. 5.11 shows the results for heavy street noise and non-stationary babble noise.

5.2.5 Role of β Value in Noise Estimation

In this subsection, we will consider the role of both β and p values for the noise estimation by using the modulation domain based perceptual considerations.

Let us first consider the role of the β value. The power laws have been utilized in past to model the nonlinear relation between the intensity of sound and its perceived loudness [131]. Since loudness is more perceptually relevant than the sound's intensity, a cost function which would consider the difference in terms of the perceived loudness would be preferable to cost functions which consider the difference in terms of the sound intensity [123]. As apposed to the power law, the exponent β of $1/3$ was suggested in [131] to perform the nonlinear transformation between intensity and perceived loudness.

As can be observed from Fig. 4.2, lower exponent β value gives smaller speech gain (sub-fig 4.2b), which should, therefore, produce more noise reduction but will, however, also introduce more speech distortion. Contrary to this, the chance of the speech distortion by using the noise estimator (sub-fig 4.2c) is reduced as because the noise gain is lowered by decreasing the exponent β value. This difference between noise and speech estimators can be seen clearly from the results plotted in the subsection 5.2.4. The role of the exponent p used in the weighted β-MMSE noise estimator is given along with the exponent p employed in the $WCOSH$ noise estimator in the subsection 5.3.3.

5.3 Weighted $COSH$ Noise Estimator

5.3.1 $WCOSH$ Gain with Large *a posteriori* SNR ($\gamma \gg 1$)

As we have seen from subsection 5.2.2 that ν_n has the asymptotic relation with γ and ξ, that clearly means that, when $\gamma \to \infty$, the ν_n also approaches to ∞, which therefore allows to use the approximation of the confluent hyper-geometric function by using [125, Eq. A1.16b]. For convenience, the noise gain function of using the $WCOSH$ estimator and the approximated confluent hyper-geometric function are repeated here as

$$G_N = \sqrt{\frac{\xi}{\gamma(\xi+1)} \frac{\Gamma\left(\frac{p+3}{2}\right)}{\Gamma\left(\frac{p+1}{2}\right)} \frac{\Phi(-\frac{p+1}{2}, 1; -\nu_n)}{\Phi(-\frac{p-1}{2}, 1; -\nu_n)}} \quad \forall \;\; p > -1, \tag{5.19}$$

and

$$\lim_{\nu_n \to \infty} \Phi(\mathrm{a}, 1; -\nu_n) \approx \frac{\nu_n^{-\mathrm{a}}}{\Gamma(1-\mathrm{a})}. \tag{5.20}$$

By using the confluent hyper-geometric function approximation for large value of $\gamma(\gg 1)$, the noise gain function given in Eq. (5.19) represents

$$\lim_{\nu_n \to \infty} G_N \approx \sqrt{\frac{\xi}{\gamma(\xi+1)} \left[\frac{\Gamma\left(\frac{p+3}{2}\right) \Gamma\left(\frac{p+1}{2}\right) (\nu_n)^{\frac{p+1}{2}}}{\Gamma\left(\frac{p+1}{2}\right)\Gamma\left(\frac{p+3}{2}\right)(\nu_n)^{\frac{p-1}{2}}} \right]}, \tag{5.21}$$

which is simplified as

$$\lim_{\nu_n \to \infty} G_N \approx \sqrt{\frac{\xi\ \nu_n}{\gamma(\xi+1)}}. \tag{5.22}$$

By letting $\nu_n = \frac{\gamma}{\xi(\xi+1)}$, the *WCOSH* based noise gain function given in (5.19) is approximated by

$$\lim_{\nu_n \to \infty} G_N \approx \frac{1}{\xi+1}. \tag{5.23}$$

From (5.15) and (5.23), it is clearly observed that both the weighted β-MMSE and *WCOSH* derived noise estimators perform as a Wiener noise gain for the larger instantaneous SNRs values.

5.3.2 *WCOSH* Speech & Noise Estimators Performance

For the experimental setup, the same stimuli were used with all varying stationarity noise signals. The WCOSH noise estimator with varying p (-1 $< p \leq$ 2) value is used. For comparison purpose, the parent WCOSH speech estimator derived in [21] is considered and re-written here, as

$$G_S = \frac{1}{\gamma} \sqrt{\frac{\nu_k\ \Gamma\left(\frac{p+3}{2}\right)}{\Gamma\left(\frac{p+1}{2}\right)} \frac{\Phi(-\frac{p+1}{2}, 1; -\nu_k)}{\Phi(-\frac{p-1}{2}, 1; -\nu_k)}} \quad \forall\ \ p > -1, \tag{5.24}$$

Note that, the difference between speech and noise estimator is that, the speech estimator uses $\nu_k = \frac{\xi\gamma}{\xi+1}$, whilst the noise estimator uses $\nu_n = \frac{\nu_k}{\xi^2} = \frac{\gamma}{\xi(\xi+1)}$.

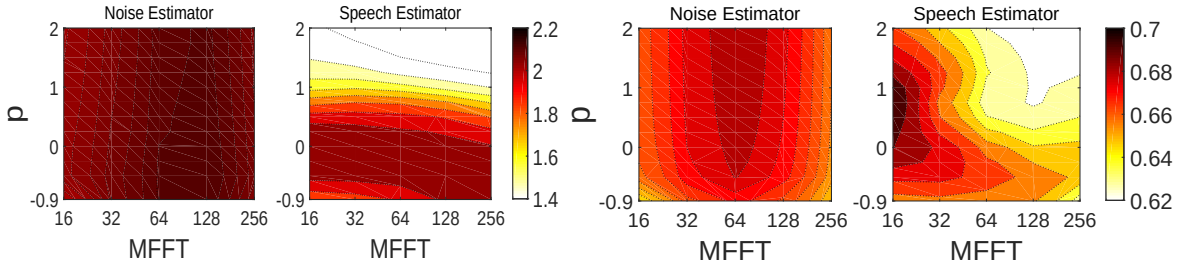
The PESQ and STOI score based results achieved by using both the WCOSH speech and noise estimators are plotted in Fig. 5.12, 5.13, 5.14, and 5.15 for stationary white noise, factory noise, heavy street noise, and non-stationary babble

noise respectively, whilst Figs. 5.16, and 5.17 represent the respective segmental SNR measures of using these noise signals.

It is observed for Fig. 5.12 that, the effect of the exponent p in estimating the noise spectral amplitudes is less as the intelligibility of the enhanced speech does not change much by changing the exponent p values. Contrary to this, the speech estimator completely depends on the selection of the appropriate p values. Moreover the range of the exponent p value for estimating the speech spectral coefficients is limited whilst, this range is larger for the noise estimation. For instance, the noise estimator provides the better PESQ score for all the exponent p values by using the modulation FFT of 16 and 32, but it is restricted to use the exponent p in between 0.25 to 0.60 in speech estimator. Similarly, the STOI score suggests the similar limitation in speech estimator whilst the noise estimator can adopt according to the exponent p value provided. On the other hand, both the modulation FFT size 16 and 32 promise better intelligibility score by using the noise estimator. However, STOI score suggests that the speech estimator has the better score than the noise estimator, but clearly it is limited to the specific values. As for as the segmental SNR is concerned, Fig. 5.16 clearly suggests that for all the negative value of the exponent p, the speech estimator gives better segmental SNR. It is also validated by the previous research conducted by Loizou in [21], where he clearly mentioned that the better performance is obtained with negative values of p in the speech estimator.

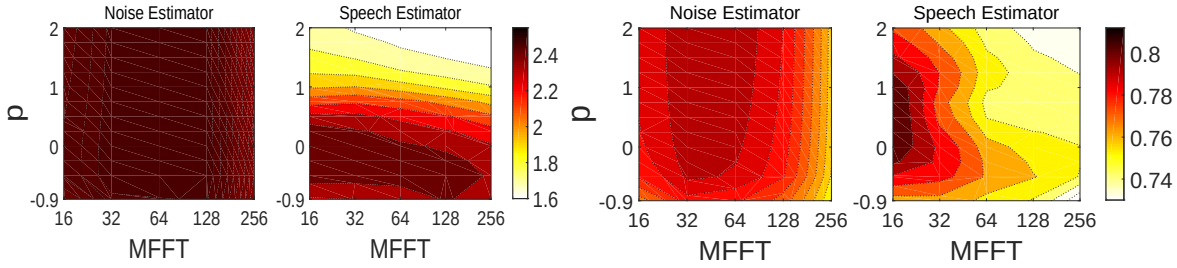
The similar performances achieved for the long-term stationary factory noise, i.e., shown in Fig. 5.13, where reducing the modulation FFT size (16 and 32) the noise estimator achieves better intelligibility scores and the SNRseg of the enhanced speech. Whilst, the speech estimator, is restricted by a limited range of the MFFT and the exponent p values.

The behavior of both speech and noise estimators toward the street and babble noise is different as what we have noticed with the stationary white noise and the factory noise. More specifically, for the heavy street noise as plotted in Fig. 5.14 and 5.17, the modulation FFT size 32 sounds good in estimating the noise, but for a particular value of the exponent p, speech estimator achieves better intelligibility. As this performance of the speech estimator may be better suited



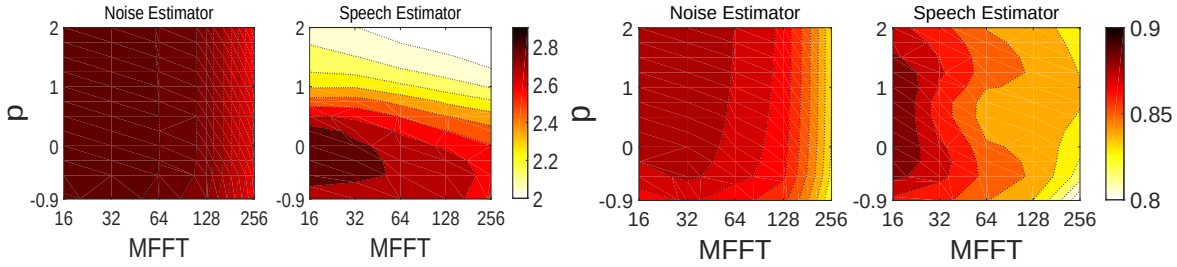
(a) PESQ (0 dB).

(b) STOI (0 dB).



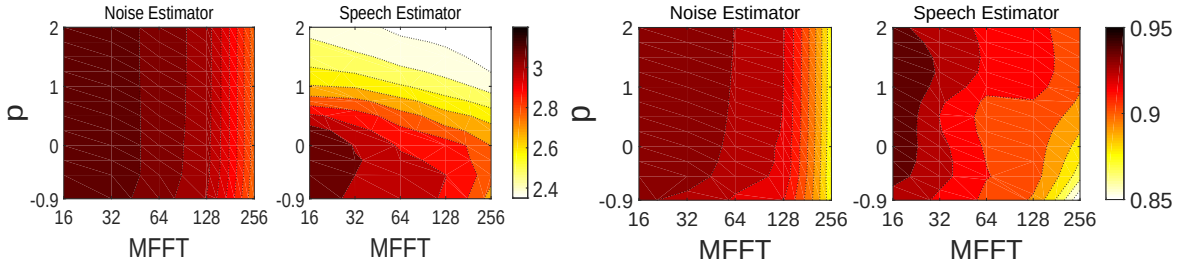
(c) PESQ (5 dB).

(d) STOI (5 dB).



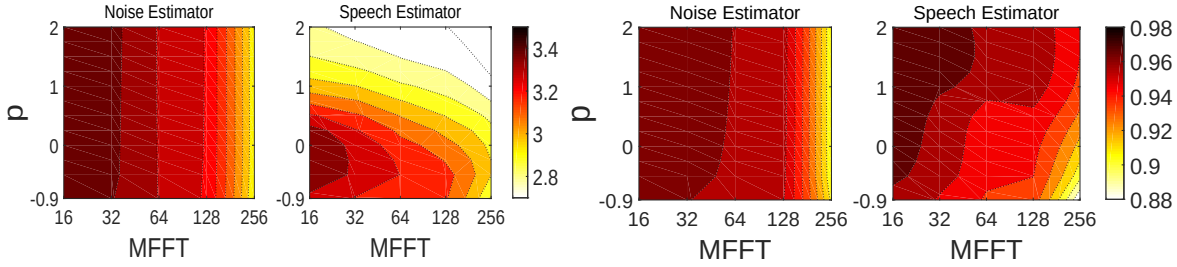
(e) PESQ (10 dB).

(f) STOI (10 dB).



(g) PESQ (15 dB).

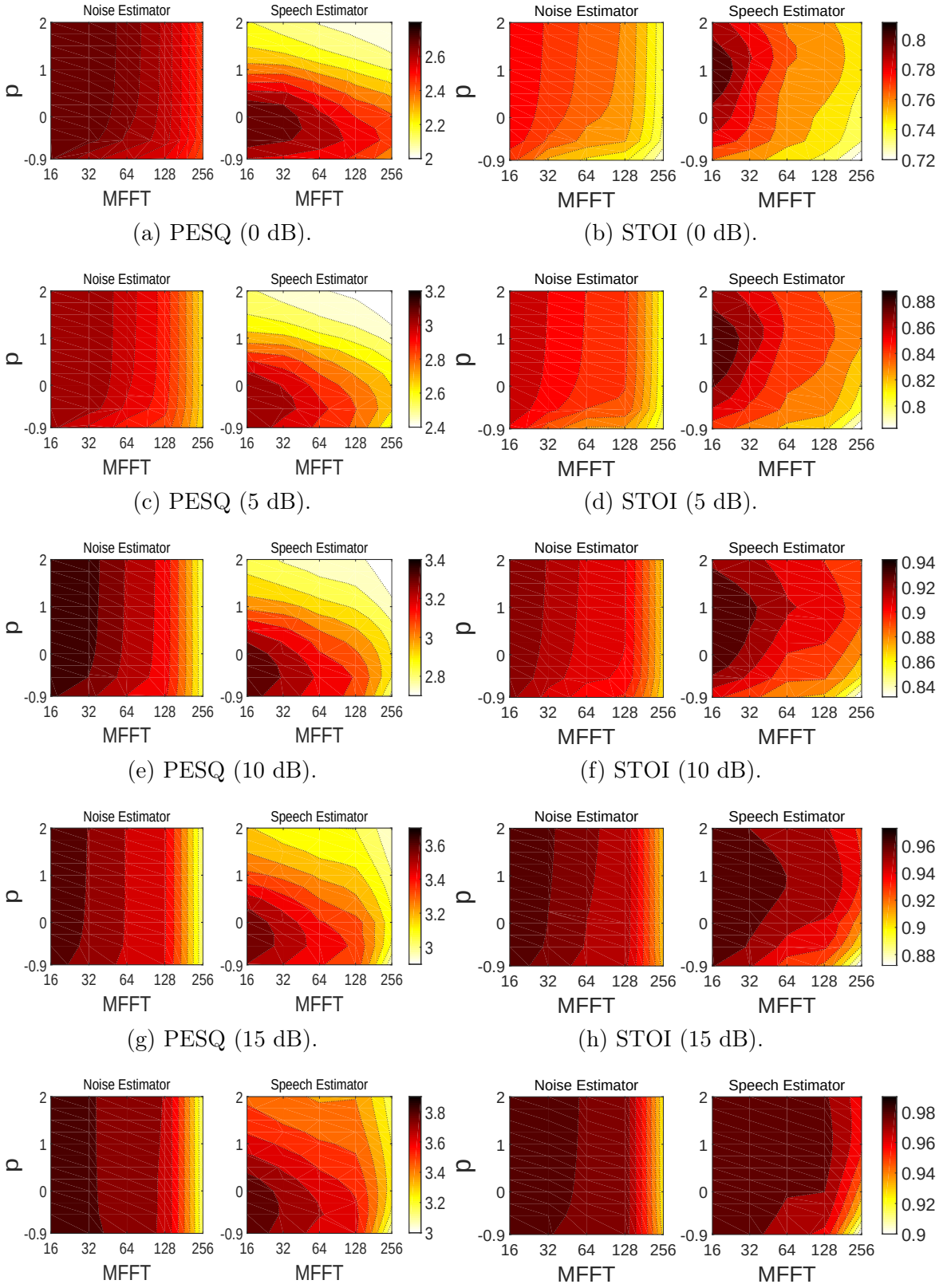
(h) STOI (15 dB).



(i) PESQ (20 dB).

(j) STOI (20 dB).

Figure 5.12: The stationary white noise based mean PESQ (a, c, e, g, i) and STOI scores (b, d, f, h, j) of enhanced speech achieved by using both the speech and noise estimators derived from weighted COSH estimator.



(a) PESQ (0 dB). (b) STOI (0 dB).

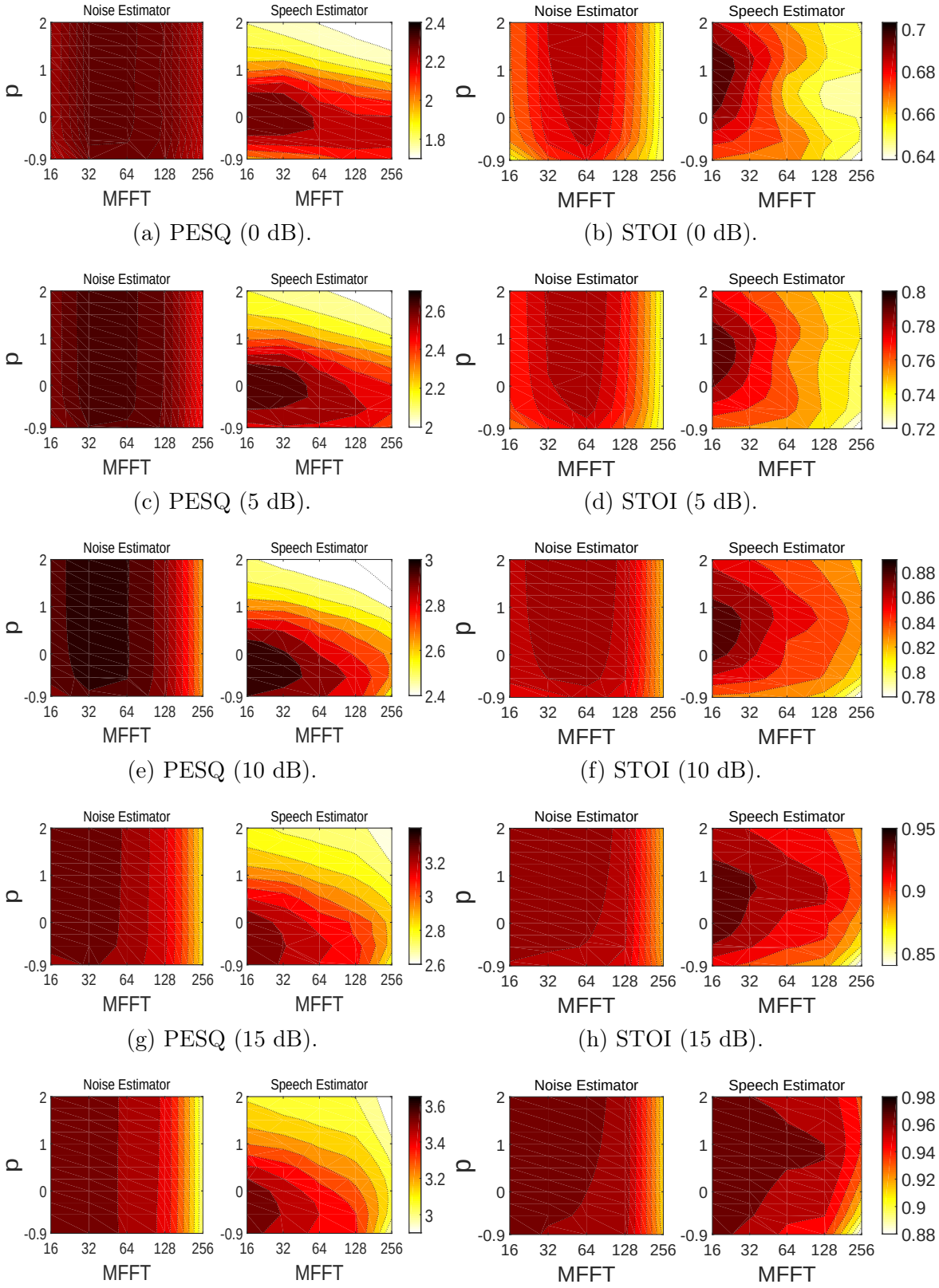
(c) PESQ (5 dB). (d) STOI (5 dB).

(e) PESQ (10 dB). (f) STOI (10 dB).

(g) PESQ (15 dB). (h) STOI (15 dB).

(i) PESQ (20 dB). (j) STOI (20 dB).

Figure 5.13: The long-term stationary factory noise based mean PESQ (a, c, e, g, i) and STOI scores (b, d, f, h, j) of enhanced speech achieved by using both the speech and noise estimators derived from weighted COSH estimator.



(a) PESQ (0 dB). (b) STOI (0 dB).

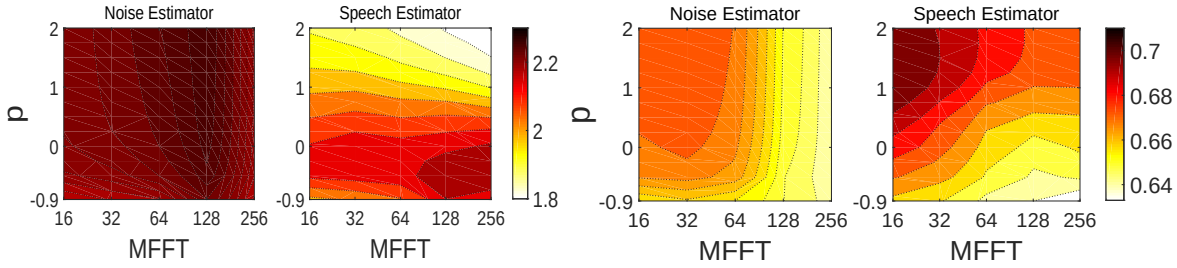
(c) PESQ (5 dB). (d) STOI (5 dB).

(e) PESQ (10 dB). (f) STOI (10 dB).

(g) PESQ (15 dB). (h) STOI (15 dB).

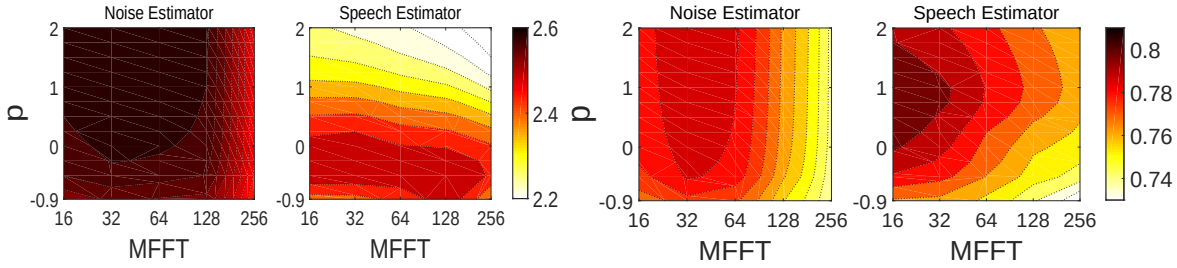
(i) PESQ (20 dB). (j) STOI (20 dB).

Figure 5.14: The heavy street noise based mean PESQ (a, c, e, g, i) and STOI scores (b, d, f, h, j) of enhanced speech achieved by using both the speech and noise estimators derived from weighted COSH estimator.



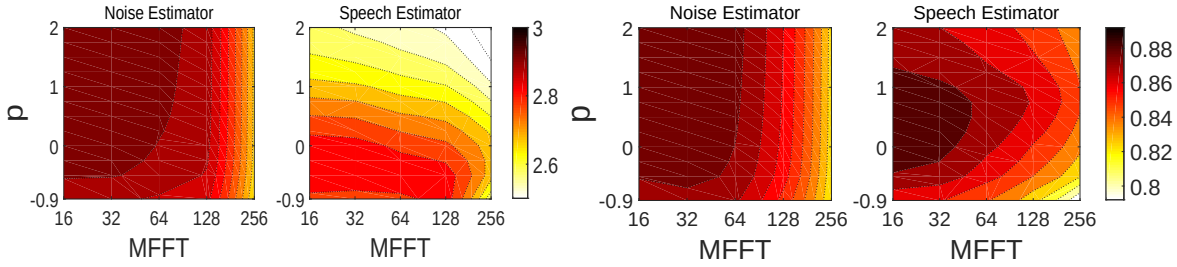
(a) PESQ (0 dB).

(b) STOI (0 dB).



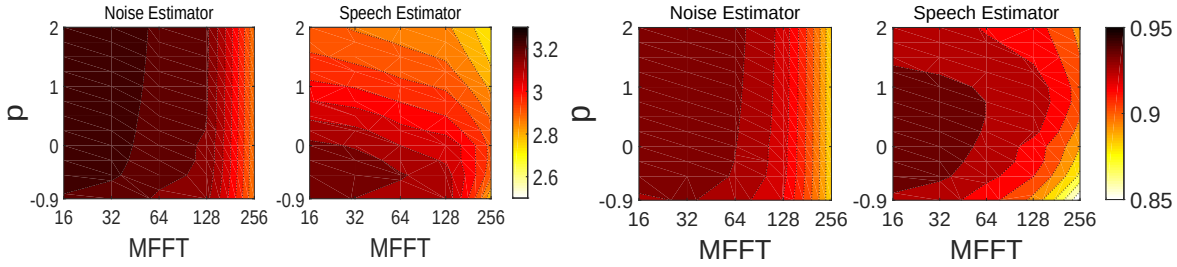
(c) PESQ (5 dB).

(d) STOI (5 dB).



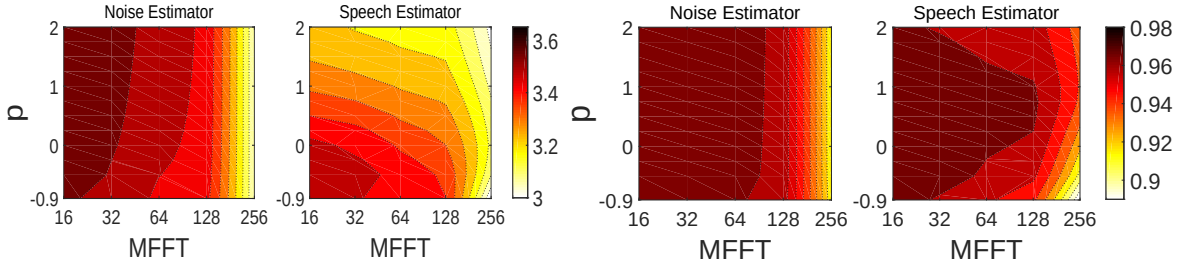
(e) PESQ (10 dB).

(f) STOI (10 dB).



(g) PESQ (15 dB).

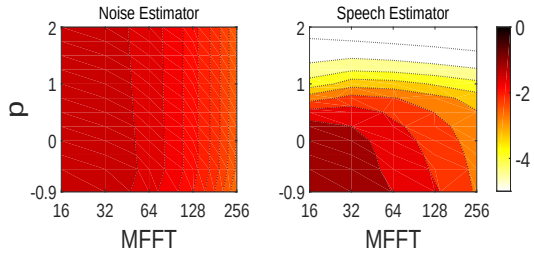
(h) STOI (15 dB).



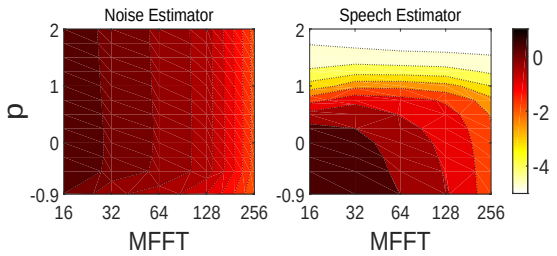
(i) PESQ (20 dB).

(j) STOI (20 dB).

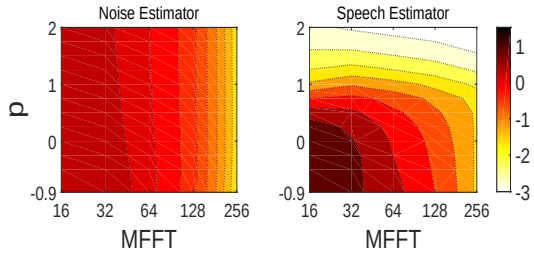
Figure 5.15: The non-stationary babble noise based mean PESQ (a, c, e, g, i) and STOI scores (b, d, f, h, j) of enhanced speech achieved by using both the speech and noise estimators derived from weighted COSH estimator.



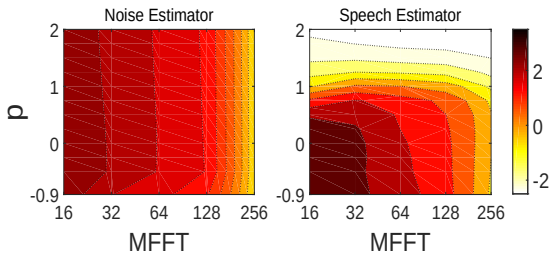
(a) 0 dB white noise.



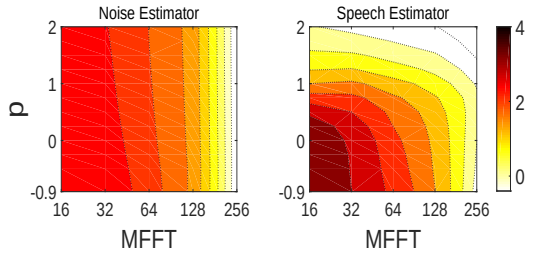
(b) 0 dB factory noise.



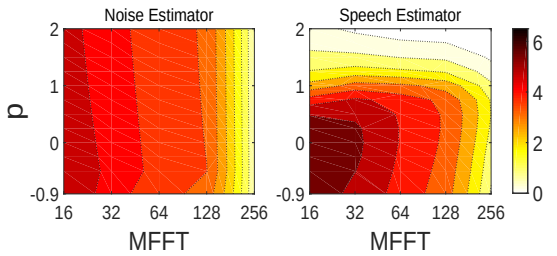
(c) 5 dB white noise.



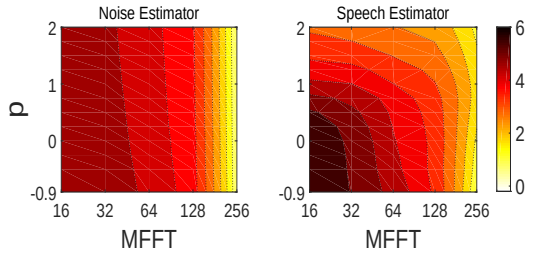
(d) 5 dB factory noise.



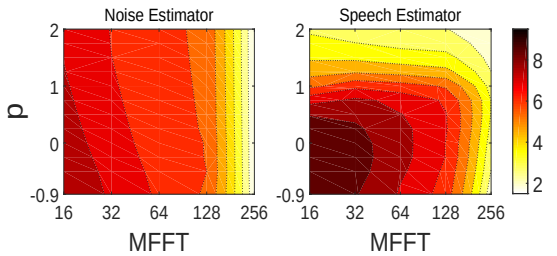
(e) 10 dB white noise.



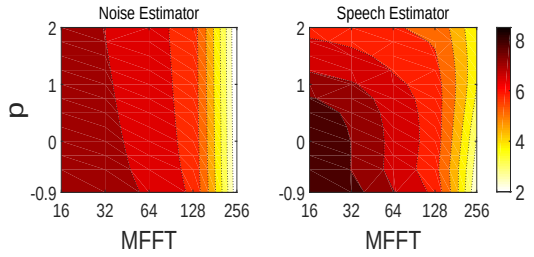
(f) 10 dB factory noise.



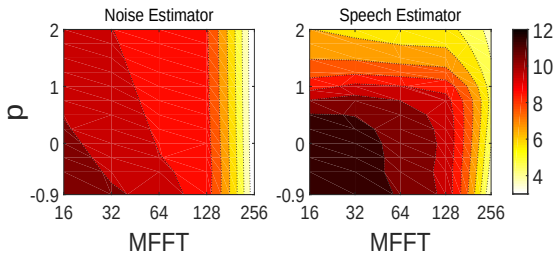
(g) 15 dB white noise.



(h) 15 dB factory noise.

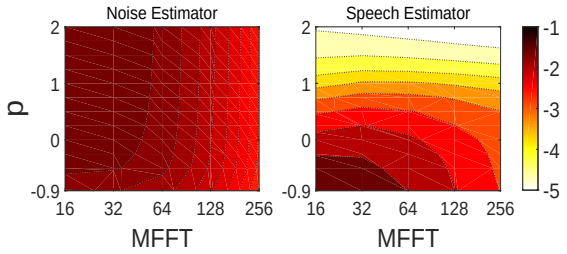


(i) 20 dB white noise.

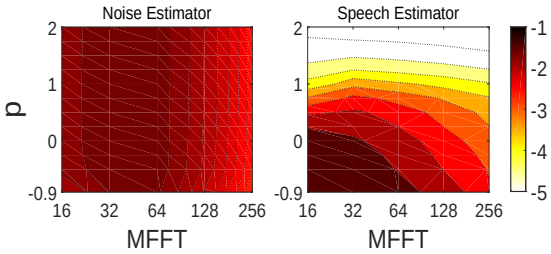


(j) 20 dB factory noise.

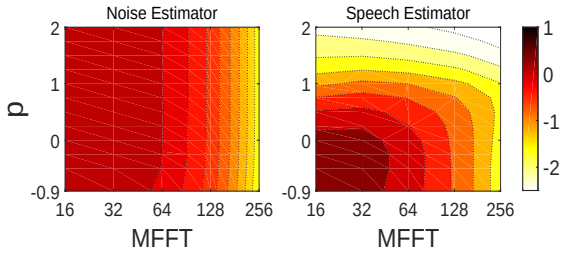
Figure 5.16: The modulation based mean SNRseg score comparison between WCOSH based proposed noise method (4.41) and the given speech estimator. The processed speech degraded by stationary white noise (a,c,e,g,i) and factory noise (b,d,f,h,j).



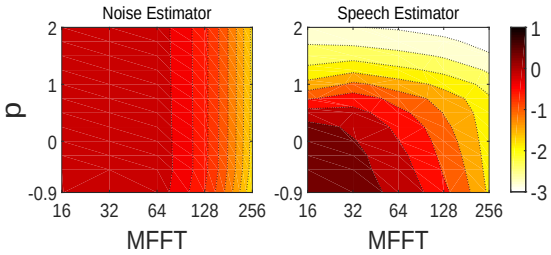
(a) 0 dB babble noise.



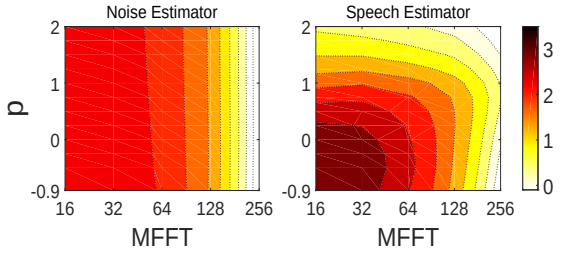
(b) 0 dB street noise.



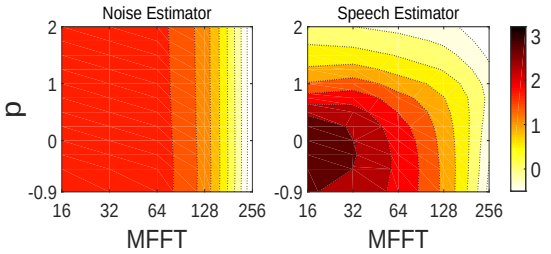
(c) 5 dB babble noise.



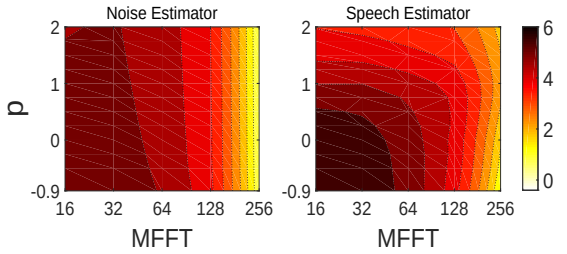
(d) 5 dB street noise.



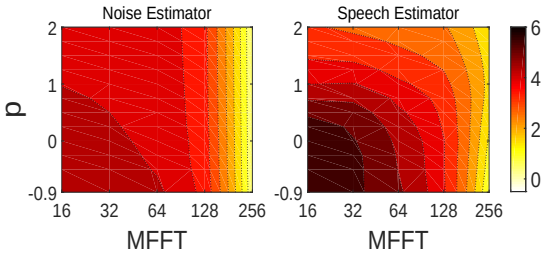
(e) 10 dB babble noise.



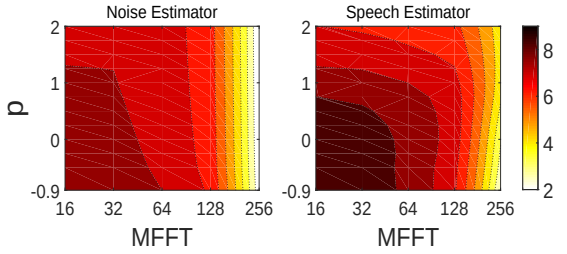
(f) 10 dB street noise.



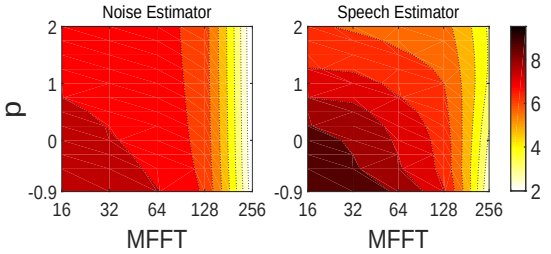
(g) 15 dB babble noise.



(h) 15 dB street noise.



(i) 20 dB babble noise.



(j) 20 dB street noise.

Figure 5.17: The modulation based mean SNRseg score comparison between WCOSH based proposed noise method (4.41) and the given speech estimator. The processed speech degraded by non-stationary babble noise (a,c,e,g,i) and heavy street noise (b,d,f,h,j).

for a particular application, but certainly it may not be the general case and therefore the applicability of this speech estimator is reduced.

5.3.3 Role of p Value in Noise Estimation

We now look at the choice of the p value that is used in both the weighted β-MMSE and WCOSH based noise estimators. The motivation for deriving the Bayesian-based noise estimators was to favor a more accurate noise estimation as noise signal is considered to be more stationary than the speech signals and therefore controlling the trade-off in noise estimation may be easier due to slow varying spectra. The numerical value p=−0.50 that controls the trade-off between the speech distortion and residual noise in the WCOSH speech estimator is suggested in [21]. It is also proposed in [21] that, the WCOSH speech estimator performs comparably with the log-MMSE estimator, but with substantially reduced residual noise. However, there is no such limitation found in the noise estimator as it works satisfactorily with all values of the exponent p. This is because the exponent used in the noise estimation controls the trade-off between the speech distortion and noise estimation more efficiently by exploiting the implicitly auditory masking effects and taking into account the fact that estimation errors near the spectral peaks are masked.

5.4 Summary

This chapter successfully investigated the performance by using various configurations of the Bayesian motivated noise estimators developed in chapter 4. Particularly, we were interested in identifying and studying the response of these noise estimators in comparison with the existing parent speech estimators in the modulation domain. Since, speech intelligibility varies over the modulation FFT (MFFT) size and frame shift (MFS) and with different types the estimator, various combination of MFFT and MFS were investigated by using the Bayesian motivated noise estimators. It is evidently found from comprehensive Experiments that 32-point MFFT achieving 50% MFS gives overall better intelligibility improvement for all time-varying noise signals.

In dealing with weighted β-MMSE noise estimator, firstly it is analytically provided that the gain function follows the well known Wiener noise gain function for high instantaneous SNR ($\gamma \gg 1$). Since exponents β and p control the performance of weighted β-MMSE noise estimator, selecting appropriate values to achieve considerable improvements in modulation domain have therefore been investigated. Results indicated that lowering the p value provides better noise estimate that results in the reduction of speech distortion. Besides that, there is not much difference is noticed while changing the exponent β values.

On the other hand, the WCOSH noise estimator has been inspected by tuning the appropriate exponent p for different modulation FFT sizes. For a given MFFT size, results achieved by using the WCOSH noise estimator suggested that the estimator yields almost similar result irrespective of different p values. However, 32-point MFFT achieves considerable improvements in the speech intelligibility. This nature of both the estimators clearly substantiate the fact that the spectral weighting in the modulation domain may be useful in estimating the noise DFT amplitudes compared to the speech because in the modulation domain the spectral peakedness is reduced that neutralizes effect of estimation error due to the spectral peaks and valleys.

$\sim * \sim$

Chapter 6

Conclusion

You cannot believe in God, until you believe in yourself.
–Swami Vivekananda.

6.1 Summary of the work

For a single-channel based noise estimation, several methods exist for removal of the additive background noise, but most of these noise methods follow the Gaussian assumption for noise, and therefore the results remain inconclusive as no single density function can represent the different real world noise. This complication can be reduced by using Bayesian estimation theory, as it minimizes the Baye's risk function, which includes a posterior probability model of the unknown parameters (given from the observation vector) and a cost-error function. The more peaked the noise pdf, the larger the estimation error will be, and as a result, the greater the influence on the outcome of the noise estimation process.

Motivated by that, the modulation based Bayesian approach to estimate both the stationary and non-stationary noise signals has been derived. Since the Gaussian assumption for all noise DFT coefficients does not necessarily hold, chapter 3 investigates the best noise distribution in both the frequency and modulation domain for the speech applications. From the investigative experiments conducted by using all time-varying noise signals, it is found that the Gamma density overall yields the least deviation from the true noise distribution. The Gamma density

is then used to derive the noise estimator by using a minimum mean square error (MMSE) based Bayesian estimator in the modulation domain. The results show that the modulation domain in noise estimation contributes to a large extent towards speech intelligibility, and consequently, the erroneous noise estimate is reduced by using the Gamma density for time-varying noise signals in the short time modulation domain. In particular, an overall improvement of the proposed scheme is registered in terms of PESQ, STOI and segmental SNR in the modulation domain compared to the acoustic domain and other state of the art methods.

As the Bayesian cost functions provide perceptually meaningful speech estimators that correlate to the human auditory system, there would be a significant attention, from a spectral stationarity point of view, in performing the noise estimation by using the same Bayesian cost functions in the modulation domain. Therefore, these perceptually meaningful Bayesian cost functions are extended for deriving the family of noise estimators in chapter 4 by considering that the noise DFT coefficients are comparably more stationary than the speech DFT coefficients and tracking the noise spectral amplitudes become easier by using these Bayesian noise estimators. Moreover, these generalized estimators represent the family of estimators and therefore provide more flexibility to choose the appropriate parameter to achieve the optimum performance in terms of the speech intelligibility.

The Chapter 5 provided the modulation based noise estimation by using all analytically derived adaptive Bayesian motivated noise estimators, which is important when dealing with the time-varying acoustic environments, and non-stationary noise signals such as street and multi-talker babble noise. The investigative result achieved by the noise estimators is compared with the existing parent speech estimators in the sort-time modulation domain. These results have led to many interesting developments for estimating noise spectral amplitudes in the modulation domain. Such as the noise estimator is more adaptive compared to the speech estimators, whilst the modulation FFT size and frame shift play most crucial role in achieving better results. The exponent β in the weighted-β-MMSE noise estimator is found to be insensitive as the estimator delivers almost similar results for all values. Contrary to this, spectral weighting (exponent p) in the

modulation domain effectively captivate the performance of the speech estimator by limiting the range of values whilst the exponent p used in both weighted-β-MMSE and WCOSH based noise estimators, on the other hand, allows the noise estimator to use a large range of values.

6.2 Future Research Directions

Since, modulation domain has apparent advantage in speech quality improvement over frequency domain, it can be potentially combined with many existing speech enhancement techniques implemented in frequency domain and further improvement can be achieved. The promising directions for future research have emerged based on the work presented in this Thesis. These are summarized briefly below:

- ***Selection of modulation Frequency based adaptive noise exponents*** β ***and*** p ***values:-*** As the family of noise estimators have been derived in Chapter 4 and implemented in Chapter 5 respectively, it would be more promising to investigate the relation of these exponents β and p such that the variation of noise statistics for each modulation frame can be better approximated and as a result, better noise estimate will be achieved. This may indeed result in still better noise tracking performance in the short-time modulation domain.

- ***Selection of Modulation Frame Length & Shift:-*** It is clearly found from Chapter 3 that the intelligibility of the speech depends on the size of the selected FFT size and frame shift. However, the problem of choosing suitable modulation FFT size and shift remain inconclusive as smaller FFT size provides higher intelligibility which is preferred in speech based applications such as hearing aid devices. Whilst, applications such as speech coding prefers better quality over intelligibility and therefore selection of modulation FFT size may differ by choosing large FFT size. Imposing such constraints restricts the applicability of the algorithms. Therefore, an adaptive system is needed which assures the effective applicability in both speech quality and intelligibility preferred devices.

- ***Intelligibility improvements over existing noise models:-*** Another important direction for future research is to study how intelligibility of enhanced speech signals can be improved with respect to existing enhancement algorithms. In general, existing noise methods improve quality in terms of noise suppression, but decrease quality in terms of speech intelligibility. A challenging direction of research would be to investigate how the decrease in intelligibility can be restricted while still obtaining good noise reduction.

- ***Incorporating multiple speech enhancement models in one system:-*** Besides further development of single-channel speech enhancement systems, it would also be interesting to investigate how multiple speech enhancement systems can cooperate in an adaptive manner to achieve better estimation. Interestingly, research on this jointly type of processing will be challenging, and might lead to a different and new view on speech enhancement and might change the insight in how to solve the speech enhancement problem.

- ***Real-time implementation of modulation domain processing:-*** The real time implementation of modulation domain processing has not been well studied yet. Although, spectral subtraction used methods are computationally inexpensive, they have less industrial applications than adaptive filtering methods due to the speech distortion introduced by spectral subtraction. One notable advantage of modulation domain processing is the speech distortion reduction, and it would have wide applications if its real time implementation can be achieved.

- ***Need for improvements in Quality & Intelligibility Assessments:*** In this thesis, different objective evaluation measures have been used to predict the quality of the speech enhanced by noise reduction algorithms. However, most of them are not really consistent in performance over a wide range of non-stationary speech and noise scenarios. Thus, another pathway for future research directions is to design an objective evaluation metric that can better predict the performance in both speech quality and intelligibility. It is also desirable to conduct future evaluation on more speech and noise databases. In addition, one of the future works for the binaural speech

enhancement algorithms are to conduct formal subjective listening tests to justify the results obtained from the objective evaluation measures.

6.3 Final Remark

The work in this thesis has led to many interesting developments in speech enhancement. Notably, modulation spectrum has a more predictive spectral variation as it reduces the spectral peakedness which helps estimator to be more effective. As a results, tracking the noise spectrum is relatively easier in the modulation domain and noise estimator successfully tracked the noise variations compared to the speech estimators by reducing the speech distortions.

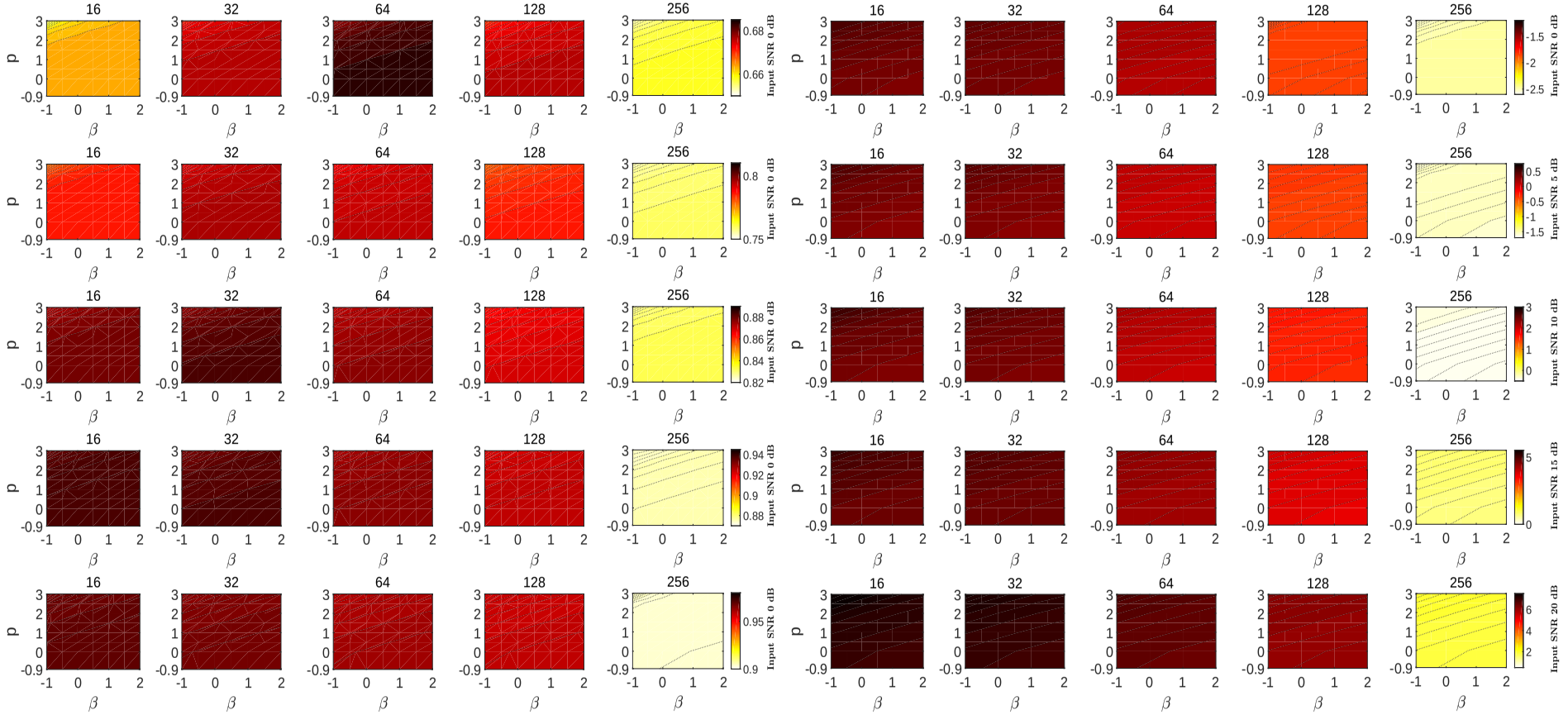
Since, the performance of the estimators depends on the noise types and SNR levels of the noisy speech, problem of estimating the time-varying noise amplitudes will, most likely be formed of many different approaches and not of one single elusive noise model. The modulation domain will, therefore, be a well-engineered alternative to the frequency domain, when dealing with the time-varying acoustic environment, and non-stationary noise signals such as street and multi-talker babble noise.

$\sim * \sim$

Appendix A

Role of MFFT Size by Varying β & p Values

A.1 Stationary White Noise Based Performance

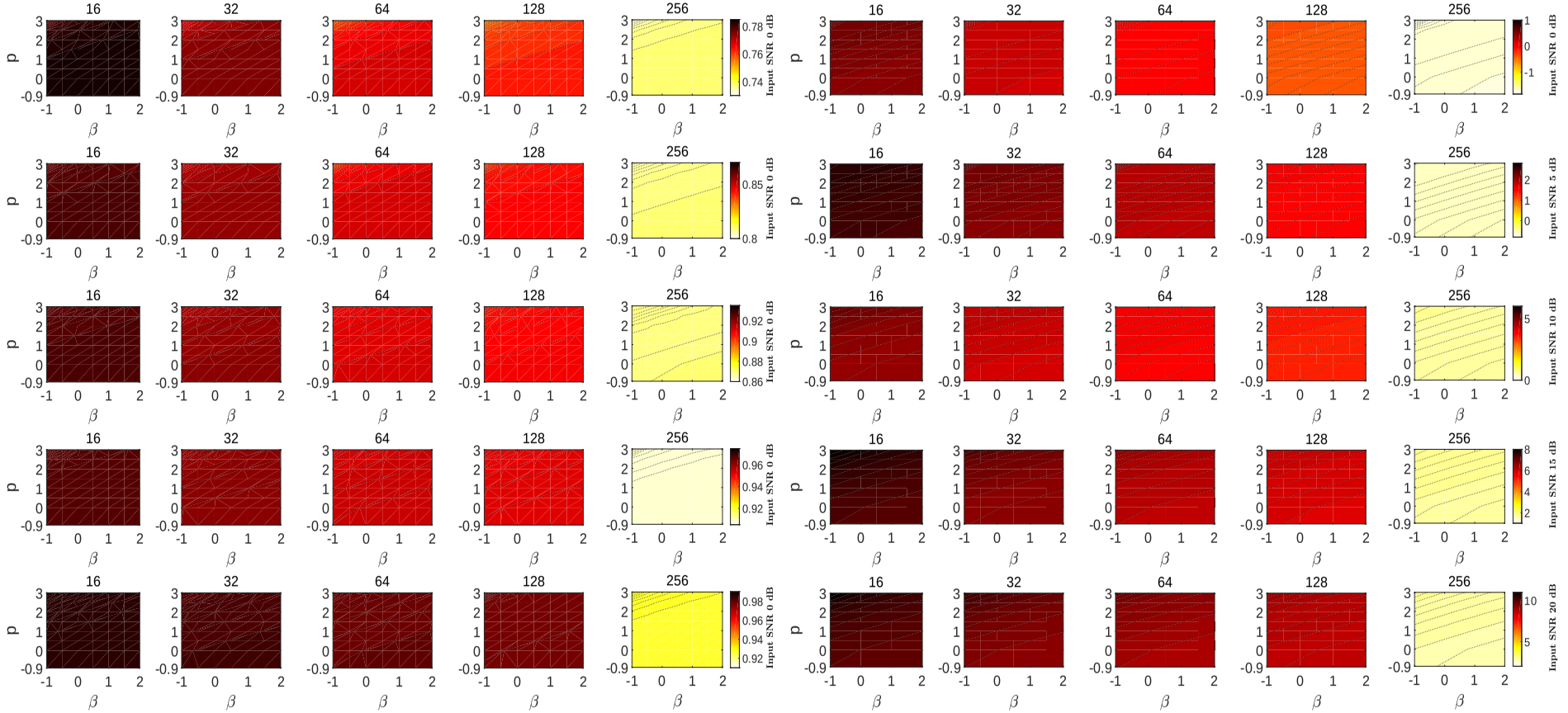


(a) The mean STOI score.

(b) The mean SNRseg.

Figure A.1: The performance for varying modulation FFT size (MFS 50%), β, and p values.

A.2 Factory Noise (Long-Term Stationary) Based Performance

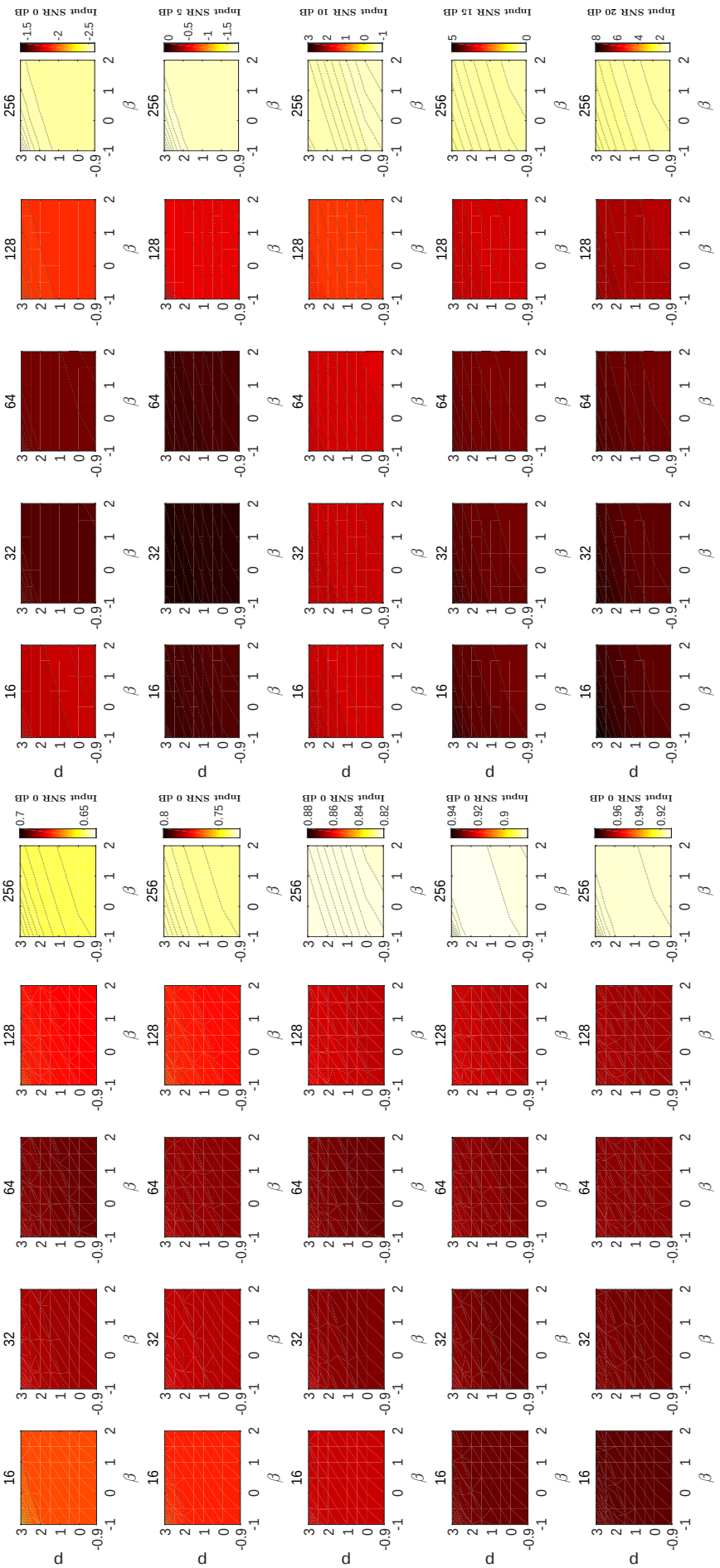


(a) The mean STOI score.

(b) The mean SNRseg.

Figure A.2: The performance for varying modulation FFT size (MFS 50%), β, and p values.

A.3 Heavy Street Noise Based Performance

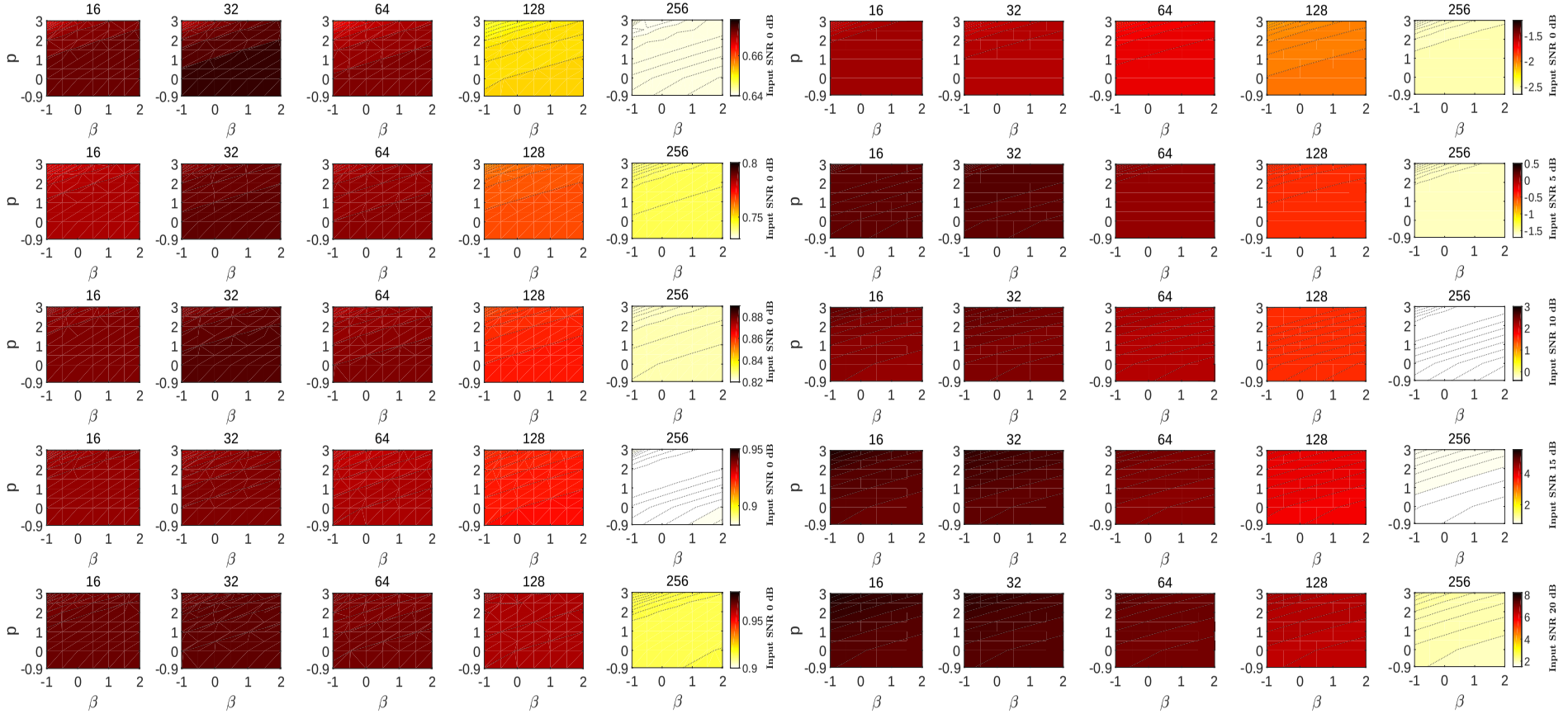


(a) The mean STOI score.

(b) The mean SNRseg.

Figure A.3: The performance for varying modulation FFT size (MFS 50%), β, and p values.

A.4 Highly Non-Stationary Babble Noise Based Performance



(a) The mean STOI score.

(b) The mean SNRseg.

Figure A.4: The performance for varying modulation FFT size (MFS 50%), β, and p values.

This Page Intentionally Left Blank.

Bibliography

[1] N. D. Degan and C. Prati, "Acoustic noise analysis and speech enhancement techniques for mobile radio applications," *Signal Processing*, vol. 15, pp. 43–56, July 1988.

[2] M.M. Goulding and J. S. Bird, "Speech enhancement for mobile telephony," *IEEE Transactions on Vehicular Technology*, vol. 39, pp. 316–326, November 1990.

[3] C. H. You, S. N. Koh and S. Rahardja, "Adaptive β-order MMSE speech enhancement application for mobile communication in a car environment," in *Fourth International Conference on Information, Communications and Signal Processing*, vol. 3, pp. 1629–1632 vol.3, Dec 2003.

[4] M. Li, H. G. McAllister, N. D. Black and T. A. Perez, "Perceptual time-frequency subtraction algorithm for noise reduction in hearing aids," *IEEE Transactions on Biomedical Engineering*, vol. 48, pp. 979–988, Sept 2001.

[5] A. Spriet, M. Moonen and J. Wouters, "Robustness analysis of multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 487–503, July 2005.

[6] T. J. Klasen, T. V. Bogaert, M. Moonen and J. Wouters, "Binaural Noise Reduction Algorithms for Hearing Aids That Preserve Interaural Time Delay Cues," *IEEE Transactions on Signal Processing*, vol. 55, pp. 1579–1585, April 2007.

[7] O. Roy and V. Martin, "Rate-Constrained Collaborative Noise Reduction for Wireless Hearing Aids," *IEEE Transactions on Signal Processing*, vol. 57, pp. 645–657, Febraury 2009.

[8] B. Widrow, "A microphone array for hearing aids," *IEEE Circuits and Systems Magazine*, vol. 1, pp. 26–32, Second 2001.

[9] A. R. Fukane and S. L. Sahare, "Enhancement of Noisy Speech Signals for Hearing Aids," in *Communication Systems and Network Technologies (CSNT), 2011 International Conference on*, pp. 490–494, June 2011.

[10] H. W. Lollmann and P. Vary, "Low Delay Noise Reduction and Dereverberation for Hearing Aids," *EURASIP Journal on Advances in Signal Processing*, pp. 1–9, April 2009.

[11] J. H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 59–71, January 1995.

[12] R. Martin and R. V. Cox, "New speech enhancement techniques for low bit rate speech coding," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 165–167, June 1999.

[13] T. Agarwal and P. Kabal, "Pre-processing of noisy speech for voice coders," in *in Proceedings IEEE Speech Coding Workshop.*, pp. 169–171, October 2002.

[14] J. H. L. Hansen and M. A. Clements, "A Constrained iterative speech enhancement with application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 39, pp. 795–805, April 1991.

[15] T. Yamada, M. Kumakura and N. Kitawaki, "Performance Estimation of Speech Recognition System Under Noise Conditions Using Objective Quality Measures and Artificial Voice," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 2006–2013, November 2006.

[16] S. Vanambathina and T. Kishore Kumar, "Speech enhancement using Bayesian estimation given a priori knowledge of clean speech phase," *Speech Communication*, vol. 77, p. 827, 2016.

[17] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 208–211, April 1979.

[18] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[19] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, pp. 1586–1604, December 1979.

[20] Y. Ephraim and D. Malah, "Speech Enhancement Using a minimum mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 1109–1121, December 1984.

[21] P. C. Loizou, "Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum," *IEEE Transactions on Speech, and Audio Processing*, vol. 13, pp. 857–869, September 2005.

[22] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, pp. 443–445, April 1985.

[23] L. Atlas, "Modulation Spectral Transforms-Application to Speech Separation and Modification," tech. rep., The Institute of Eclectronics, Information and Communication Engineers, June 2003.

[24] J. Thompson and L. Atlas, "A non-uniform modulation transform for audio coding with increased time resolution," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, pp. 397–400, April 2003.

[25] S. M. Schimmel and L. E. Atlas, "Analysis of signal reconstruction after modulation filtering," *Proc. of SPIE*, vol. 5910, 2005.

[26] S. M. Schimmel , L. E. Atlas and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 605–608, April 2007.

[27] K. Paliwal, W. Kamil and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, pp. 450–475, May 2010.

[28] S. M. Schimmel, *Theory of Modulation Frequency Analysis and Modulation Filtering, with Applications to Hearing Devices*. PhD thesis, University of Washington, 2007.

[29] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction.," *Journal of the Acoustical Society of America*, vol. 95, pp. 1593–1602, 1994.

[30] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," *International Conference on Spoken Language Processing (ICSLP)*, vol. 4, pp. 2490–2493, October 1996.

[31] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *The Journal of the Acoustical Society of America*, vol. 102, pp. 2892–2905, August 1997.

[32] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," *The Journal of the Acoustical Society of America*, vol. 102, pp. 2906–2919, August 1997.

[33] T. Kinnnunen, "Joint acoustic-modulation frequency for speaker recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 665–668, 2006.

[34] W. Kamil and P. C. Loizou, "Channel selection in the modulation domain for improved speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 131, pp. 2904–2913, April 2012.

[35] R. Drullman, J. Festen and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *Journal of the Acoustical Society of America*, vol. 95, p. 26702680, May 1994.

[36] S. M. Schimmel, K. R. Fitz and L. E. Atlas , "Frequency Reassignment for Coherent Modulation Filtering," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, p. V, May 2006.

[37] K. Paliwal, B. Schwerin and W. Kamil, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Communication*, vol. 54, pp. 282–305, February 2012.

[38] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 137–145, April 1980.

[39] Y. Rongshan, "A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4421–4424, April 2009.

[40] R. Martin, "Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 845–856, September 2005.

[41] T.H. Dat, K. Takeda, and F. Itakura, "Generalized Gamma modeling of speech and its online estimation for speech enhancement," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1520–6149, March 2005.

[42] R. C. Hendriks and R. Heusdens, "On linear versus non-linear magnitude-DFT estimators and the influence of super-Gaussian speech priors," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4750–4753, March 2010.

[43] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*. John Wiley & Sons Ltd., third ed., 2006.

[44] J. S. Garofolo, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM," *National Institute of Standards and Technology (NIST)*, pp. 1–78, October 1990.

[45] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–253, 1993.

[46] P. C. Loizou, *Speech Enhancement Theory and Practice*. CRC Press, second ed., 2011.

[47] Yi Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 229–238, January 2008.

[48] A. Rix, J Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ), a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 749–752., May 2001.

[49] John L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," *Proceedings of the International Conference on Speech and Language Processing*, pp. 2819–2822, 1998.

[50] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech,"

IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, pp. 2125–2136, February 2011.

[51] P. J. Wolfe and S. J. Godsill, "Efficient Alternatives to the Ephraim and Malah Suppression rule for Audio Signal Enhancement," *European Association for Signal Processing (EURASIP)*, vol. 10, pp. 1043–1051, 2003.

[52] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, pp. 45–57, February 1991.

[53] Y. Ephraim and H. L. Van Tree, "A Signal Subspace Approach for Speech Enhancement," *IEEE Transactions on Speech, and Audio Processing*, vol. 3, pp. 251–266, July 1995.

[54] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 12, pp. 177–180, 1987.

[55] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 373–385, July 1998.

[56] J. Sohn and N. Kim, "A Statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, January 1999.

[57] Y. D. Cho, K. AI-Naimi and A. Kondoz, "Improved statistical voice activity detection based on a smoothed statistical likelihood ratio," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 737–740, 2001.

[58] A. Davis, S. Nordholm and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactions on Audio Speech and Language Processing*, vol. 14, pp. 412–424, March 2006.

[59] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," *European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1513–1516, 1995.

[60] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech, and Audio Processing*, vol. 9, pp. 504–512, July 2001.

[61] R. Martin, "Spectral subtraction based on minimum statistics," *European Signal Processing Conference (EUSIPCO)*, pp. 1182–1185, September 1994.

[62] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, pp. 12–15, January 2002.

[63] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech, and Audio Processing*, vol. 11, pp. 466–475, September 2003.

[64] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, pp. 220–231, 2006.

[65] R. C. Hendriks, R. Heusdens and J. Jensen, "MMSE based noise PSD tracking with low complexity," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4266–4269, March 2010.

[66] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1383–1393, December 2011.

[67] P. C. Yong, S. Nordholm and H. Dam, "Noise Estimation Based on Soft Decisions and Conditional Smoothing for Speech Enhancement," *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–4, September 2012.

[68] J. B. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, pp. 235–238, June 1977.

[69] J. B. Allen & L. R. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," *Proceedings of the IEEE*, vol. 65, pp. 1558–1564, November 1977.

[70] D. Griffin and J. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 236–243, April 1984.

[71] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, Upper Saddle River, NJ., 2002.

[72] A. V. Oppenheim, J. S. Lim, G. Kopec, and S.c. Pohlig, "Phase in speech and pictures," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 632–637, April 1979.

[73] A. V. Oppenheim and J.S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, pp. 529–541, May 1981.

[74] D. Wang and J.S. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, pp. 679 – 681, August 1982.

[75] P. Vary, "Noise Suppression by Spectral Magnitude Estimation-Mechanism and theoretical limits," *Signal Processing*, vol. 8, pp. 387–400, July 1985.

[76] L. Liu, H. Jialong, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 22, pp. 403–417, September 1997.

[77] K. Wojcicki, and K. Paliwal, "Importance of the Dynamic Range of an Analysis Windowfunction for Phase-Only and Magnitude-Only Reconstruction of Speech," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 729–732, April 2007.

[78] L. D. Alsteris and K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, pp. 727–736, June 2006.

[79] T. Gerkmann, K. Martin and R. Robert, "Phase estimation in speech enhancement-Unimportant, important, or impossible?," *Convention of Electrical and Electronics Engineers*, vol. 27, pp. 1–5, November 2012.

[80] K. Paliwal, "Usefulness of phase in speech processing," *Proc. IPSJ Spoken Language Processing Workshop*, pp. 1–6, 2003. Gifu, Japan.

[81] P. C. Loizou and G. Kim, "Reasons why Current Speech-Enhancement Algorithms do not Improve Speech Intelligibility and Suggested Solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 47–56, January 2011.

[82] P. Pollak, P. Sovka, and J. Uhlir, "The Noise Supression System for a Car," *European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1073–1076, September 1993.

[83] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, vol. 54, pp. 297–315, Febraury 1975.

[84] A. L. Zadeh, "Frequency analysis of variable networks," *Proceedings of the IRE*, vol. 38, pp. 291–299, March 1950.

[85] L. Atlas, and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 668–675, February 2003.

[86] N. F. Viemeister, "Temporal Modulation Transfer Functions for Audition," *Journal of the Acoustical Society of America*, vol. 53, no. 1, pp. 312–312, 1973.

[87] N. F. Viemeister, "Modulation thresholds and temporal modulation transfer functions," *The Journal of the Acoustical Society of America*, vol. 60, no. S1, pp. S117–S117, 1976.

[88] N. F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *The Journal of the Acoustical Society of America*, vol. 66, no. 5, pp. 1364–1380, 1979.

[89] S. P. Bacon and D. W. Grantham, "Modulation masking: Effects of modulation frequency, depth, and phase," *The Journal of the Acoustical Society of America*, vol. 85, pp. 2575–2580, March 1989.

[90] T. Dau, J. Verhey, and A. Kohlrausch, "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers," *The Journal of the Acoustical Society of America*, vol. 106, pp. 2752–2760, July 1999.

[91] T. Houtgast, "Frequency selectivity in amplitude-modulation detection," *The Journal of the Acoustical Society of America*, vol. 85, pp. 1676–1680, November 1988.

[92] R. Drullman, J. Festen and R. Plomp, "Effect of temporal envelope smearing on speech reception," *Journal of the Acoustical Society of America*, vol. 95, pp. 1053–1064, Febraury 1994.

[93] H. Hermansky, "The modulation spectrum in the automatic recognition of speech," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 140–147, December 1997.

[94] A. Sek and B. C. Moore, "Mechanisms of modulation gap detection," *The Journal of the Acoustical Society of America*, vol. 111, pp. 2783–2792, June 2002.

[95] L. Atlas, Li Qin and J. Thompson, "Homomorphic modulation spectra," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 761–764, May 2004.

[96] M.S. Vinton and L. E. Atlas, "Scalable and progressive audio codec," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, pp. 3277–3280, May 2001.

[97] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, October 1994.

[98] C. Nadeu, P. Leal, and B. H. Juang, "Filtering the time sequences of spectral parameters for speech recognition," *Speech Communication*, vol. 22, pp. 315–332, September 1997.

[99] X. Xiao, E. S. Chng, and L. Haizhou, "Normalization of the Speech Modulation Spectra for Robust Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 1662 –1674, November 2008.

[100] X. Lu, S. Matsuda, M. Unoki, and S. Nakamura, "Temporal contrast normalization and edge-preserved smoothing of temporal modulation structures of speech for robust speech recognition," *Speech Communication*, vol. 52, pp. 1–11, January 2010.

[101] Yi Zhang and Y. Zhao, "Real and imaginary modulation spectral subtraction for speech enhancement," *Speech Communication*, vol. 55, pp. 509–522, May 2013.

[102] B. Schwerin and K. Paliwal, "Speech enhancement using STFT of real and imaginary parts of modulation signals," *Australasian Speech Science and Technology Association (ASSTA)*, pp. 25–28, December 2012.

[103] B. Schwerin and K. Paliwal, "Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement," *Speech Communication*, vol. 58, pp. 49 – 68, March 2014.

[104] Y. Wang, *Speech Enhancement in the Modulation Domain*. PhD thesis, Imperial College London, August 2015.

[105] V. Mani, "Speech Enhancement in Modulation Domain using Codebook-based Speech and Noise Estimation," MEng Thesis, McGill University, Montreal, Canada, February 2016.

[106] Y. Wang and M. Brookes, "Model-Based Speech Enhancement in the Modulation Domain," *IEEE Transactions on Audio Speech and Language Processing*, july 2017.

[107] R. Martin, "Speech Enhancement using MMSE short time spectral Estimation with Gamma Distributed Speech Prior," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 253–256, 2002.

[108] B. Chen and P. C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech Communication*, vol. 49, pp. 134–143, February 2007.

[109] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 1110–1126, January 2005.

[110] D. N. Joanes and C. A. Gill, "Comparing measures of sample skewness and kurtosis," *Journal of the Royal Statistical Society (Series D): The Statistician*, vol. 47, no. 1, p. 183189, 1998.

[111] H. Hermansky, E. Wan and C. Avendano, "Speech enhancement based on temporal processing," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 405–408, May 1995.

[112] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using Laplacian speech priors," *International Workshop on Acoustic, Echo, and Noise Control (IWAENC)*, pp. 87–90, September 2003.

[113] F. J. Massey, "The Kolmogorov-Smirnov Test for Goodness-of-Fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[114] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formula, Graphs, and Mathematical Tables*. Dover Publications, Inc., New York, 9th ed., November 1970.

[115] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products*. Academic Press, 7th ed., 2007.

[116] F. W. Olver, D. W. Lozier, R. F. Boisvert and C. W. Clark, ed., *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[117] K. Krishnamoorthy, *Handbook of Statistical Distributions with Applications*. Chapman & Hall/CRC, Taylor & Francis Group, 2006.

[118] P. C. Yong, S. Nordholm and H. H. Dam, "Trade-off evaluation for speech enhancement algorithms with respect to the *a priori* SNR estimation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4657–4660, 2012.

[119] P. C. Yong, S. Nordholm and H. H. Dam, "Optimization and evaluation of sigmoid function with a priori SNR estimate for real-time speech enhancement," *Speech Communication*, vol. 55, pp. 358–376, February 2013.

[120] B. Hanson and T. Applebaum, "Subband or cepstral domain filtering for recognition of lombard and channel-distorted speech," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 79–82, April 1993.

[121] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in persuit of an invariant representation of speech," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, pp. 1647–1650, April 1997.

[122] T. Esch and P. Vary, "Efficient Musical Noise Suppression for Speech Enhancement System," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4409–4412, April 2009.

[123] E. Plourde and B. Champagne, "Generalized Bayesian Estimators of the Spectral Amplitude for Speech Enhancement," *IEEE Signal Processing Letters*, vol. 16, pp. 485–488, March 2009.

[124] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, pp. 588–601, July 2006.

[125] D. Middleton, *An Introduction to Statistical Communication Theory*. Wiley-IEEE Press, April 1996.

[126] O. Cappe, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Transactions on speech, and audio Processing*, vol. 2, pp. 345–349, April 1994.

[127] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 475–486, July 2005.

[128] E. Plourde and B. Champagne, "Perceptually based speech enhancement using the weighted β-SA estimator," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4193–4196, April 2008.

[129] F. Itakura and S. Saito, "An analysis-synthesis telephony based on maximum likelihood method," *6th International Con. on Acoustics*, pp. 17–20, August 1968.

[130] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 380–391, October 1976.

[131] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, April 1990.