UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

INSTITUTE OF SOUND AND VIBRATION RESEARCH

Bayesian Algorithms for Speech Enhancement

by

I. Andrianakis

Thesis for the degree of Doctor of Philosophy

November 2007

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

INSTITUTE OF SOUND AND VIBRATION RESEARCH

Doctor of Philosophy

BAYESIAN ALGORITHMS FOR SPEECH ENHANCEMENT

by Ioannis Andrianakis

The portability of modern voice processing devices allows them to be used in environments where background noise conditions can be adverse. Background noise can deteriorate the quality of speech transmitted through such devices, but speech enhancement algorithms can ameliorate this degradation to some extent. The development of speech enhancement algorithms that improve the quality of noisy speech is the aim of this thesis, which consists of three main parts.

In the first part, we propose a framework of algorithms that estimate the clean speech Short Time Fourier Transform (STFT) coefficients. The algorithms are derived from the Bayesian theory of estimation and can be grouped according to i) the STFT representation they estimate ii) the estimator they apply and iii) the speech prior density they assume. Apart from the introduction of algorithms that surpass the performance of similar algorithms that exist in the literature, the compilation of the above framework offers insight on the effect and relative importance of the different components of the algorithms (e.g. prior, estimator) to the quality of the enhanced speech.

In the second part of this thesis, we develop methods for the estimation of the power of time varying noise. The main outcome is a method that exploits some similarities between the distribution of the noisy speech spectral amplitude coefficients within a single frequency bin, and the corresponding distribution of the corrupting noise. The above similarities allow the extraction of samples that are more likely to correspond to noise, from a window of past spectral amplitude observations. The extracted samples are then used to produce an estimate of the noise power.

In the final part of this thesis, we are concerned with the incorporation of the time and frequency dependencies of speech signals in our estimation model. The theoretical framework on which the modelling is based is provided by Markov Random Fields (MRF's). Initially, we develop a MAP estimator of speech based on the Gaussian MRF prior. In the following, we introduce the Chi MRF, which is employed in the development of an improved speech estimator. Finally, the performance of fixed and adaptive schemes for the estimation of the MRF parameters is investigated.

Acknowledgements

The pursuit of a PhD degree is a long and winding road that can go through frustration and despair, but can also offer the unique pleasure of discovery and a sense of achievement. I was lucky enough to have as a mentor and companion in this trip, a person whose unique insights were a constant source of inspiration for my 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 Paul White. Thanks so much, Paul!

I also want to extend my gratitude to Steve Elliott 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 Simon Godsill for transforming my viva from a terrifying-to-be experience, to a stimulating discussion with a very knowledgeable person and for awarding me the degree at the end of it!

Special thanks go to all the people involved in the EPSRC project, part of which was my PhD. Saeed Vaseghi, Esfandiar Zavarehei, Qin Yan, Ben Milner and Jonathan Darch, thank you for the creative atmosphere during our meetings and for all your constructive input throughout our collaboration.

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, during my stay at the ISVR. I particularly want to thank Matt Jones, Paulo Goncalves, Alessandro Beda and especially Cristobal 'Chispi' Gonzalez Diaz for making the 'office' a much more fun place to be and life in Southampton a lot more interesting. An extra special thanks also goes to Timos Papadopoulos, for all the helpful and motivating discussions and for being such a great friend to have around.

Finally, I want to thank my parents for all their love and support during these years I've been abroad. I know they have missed me a lot, and the same is also very true for me. I dedicate this thesis to them.

Contents

DECLARATION OF AUTHORSHIP

I, Ioannis Andrianakis, declare that the thesis entitled Bayesian Estimators for Speech Enhancement and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- This work was done wholly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as:
  - I. Andrianakis and P. R. White, *"Bayesian algorithms for speech enhancement"* ISVR Technical Report, No 305, Jan. 2006.
  - I. Andrianakis and P. R. White, *"MMSE speech spectral amplitude estimators with Chi and Gamma speech priors"* in International Conference on Acoustics, Speech and Signal Processing (ICASSP-06), vol. 3, pp 1068-1071 May 2006.
  - I. Andrianakis and P. R. White, *"Noise estimation based on matching the moments of the Rayleigh distribution for speech enhancement"* in Hellenic Institute of Acoustics 2006, Heraklion, Greece, Sep. 2006.
  - I. Andrianakis and P. R. White, *"On the application of Markov Random Fields to speech enhancement"* in Proc. 7th IMA Int. Conf. Mathematics in Signal Processing, Cirencester, UK, Dec. 2006.

Signed

Date

Acronyms

ACMRF	Adaptive Chi MRF
AMT	Auditory Masking Threshold
AR	Auto Regressive
CMRF	Chi MRF
DD	Decision Directed
EM	Estimation Maximisation
GMM	Gaussian Mixture Models
GMRF	Gaussian MRF
HMM	Hidden Markov Models
ICM	Iterated Conditional Modes
Im	Imaginary part
KL	Kullback Leibler
KLT	Karhunen Loeve Transform
LPC	Linear Prediction Coefficients
MAP	Maximum A Posteriori
MinS	Minimum Statistics
MMSE	Minimum Mean Square Error
MOS	Mean Opinion Score
MRF	Markov Random Field
pdf	probability density function
PESQ	Perceptual Evaluation of Speech Quality
r.v.	Random variable
Re	Real part
RMM	Rayleigh Moment Matching
SegSNR	Segmental SNR
SNR	Signal to Noise Ratio
STFT	Short Time Fourier Transform
SVD	Singular Value Decomposition
VAD	Voice Activity Detector

Mathematical notations

Chapter 2

$\hat{.}$	Estimate
$\mathbb{C}$	Cost function
e_s	Error
$P(x)$	Probability distribution function
$p(x)$	Pdf of x
$p(x\|y)$	Pdf of x conditional on y
$\mathbb{R}$	Risk

Chapter 3

A	Speech STFT amplitude
a	Shape parameter
B	Noise STFT amplitude
$\Gamma(.)$	Gamma function
γ	A posteriori SNR
γ_2	A posteriori SNR (DFT algorithms)
$D_.(.)$	Parabolic cylinder function
$\mathrm{E}[x]$	Expectation of x
$\mathrm{E}[x\|y]$	Conditional expectation of x given y
${}_1F_1(.,.,.)$	Confluent hypergeometric function
$H(.)$	Entropy
h	STFT window
θ	Scale parameter
$I(.,.)$	Mutual information
$\mathfrak{I}_G()$	MS1G integral
$\mathfrak{I}_L()$	MS1L integral
$I_0(.)$	Modified Bessel function of the first kind
i	Discrete index
$\mathbf{i}$	$\sqrt{-1}$
J	STFT window shift
K	STFT window length
k	Frequency bin index
L	Number of time frames
l	Time frame index
m	Discrete index
n	Noise
$\mathbf{N}$	Noise STFT

N_{Re}, N	Real part of $\mathbf{N}$
N_{Im}	Imaginary part of $\mathbf{N}$
ξ	A priori SNR
R	Noisy speech STFT amplitude
$\Re$	Real numbers
s	Speech
$\mathbf{S}$	Speech STFT
S_{Re}, S	Real part of $\mathbf{S}$
S_{Im}	Imaginary part of $\mathbf{S}$
σ^2	Variance of Gaussian distribution
$U(.,.)$	Symmetric uncertainty coefficient
ϕ	Speech phase
x	Noisy speech
$\mathbf{X}$	Noisy speech STFT
X_{Re}, X	Real part of $\mathbf{X}$
X_{Im}	Imaginary part of $\mathbf{X}$
ψ	Noisy speech phase
ω	Noise phase
Chapter 4	
α	DD method smoothing parameter
κ_1	Kurtosis of amplitude
κ_2	Kurtosis of Re and Im parts
N_{bin}	Number of histogram bins (KL divergence)
$p_d(m)$	Data pdf calculated from histogram (KL divergence)
$p_s(m)$	Prior evaluated on histogram bins (KL divergence)
Chapter 5	
λ	Moment smoothing parameter (adaptive estimation of a)
$N_{\mathcal{X}}$	Number of histogram bins in chi square test
O_i	Occurrences
$\bar{O}_i$	Expected occurrences
$\mathcal{X}^2$	Chi square statistic
Chapter 6	
α_p	Noisy speech smoothing parameter
α_d	Noise smoothing parameter
D	Maximum length of $\mathcal{Q}$
d	Discrete index
d_m	$\arg\min_d(\lvert r_m(d)\rvert)$
H_0	Speech absence hypothesis
H_1	Speech presence hypothesis
H_1^c	Conditional speech presence hypothesis
M	Number of Gaussian mixtures
μ	Mean of Gaussian distribution

$\mathcal{N}$	Gaussian distribution
$\mathcal{P}$	Estimate of $\mathrm{E}[\lvert\mathbf{N}\rvert^2]$
$\mathcal{Q}$	Vector of past noisy speech spectral amplitude values
$\mathcal{R}$	Smoothed noisy speech power spectrum
$\mathcal{R}_{\min}$	Minimum of $\mathcal{R}$
r_m	RMM criterion
w_m	Weight of Gaussian mixtures

Chapter 7

b_{ij}	Neighbours' weights
$\mathcal{C}$	Set of indices that define cliques
C	Set of indices that correspond to pairs of neighbours
G	Joint MRF density matrix
j	Discrete index
k_{f_0}	Frequency bin that corresponds the the pitch frequency
$\mathrm{n}(i)$	Neighbours of the i^{th} r.v. in an MRF
ξ^l	Local a priori SNR
ξ^g	Global a priori SNR
q	Number of r.v.'s in an MRF
Q	Set of indices of MRF r.v.'s
ρ_{ij}	Cross a posteriori SNR
$\mathcal{S}$	Space of r.v.'s in an MRF
w_{ij}	Constant that controls the interaction between the neighbours
$\Psi_{\mathcal{C}}(x_{\mathcal{C}})$	Arbitrary functions (Hammersley-Clifford theorem)

Chapter 1

Introduction

The continuous evolution of computers and digital systems has led to the widespread use of voice capturing and processing devices (e.g. mobile phones, hearing aids etc.). The portability of such devices enables them to be deployed in environments where background noise conditions can be adverse. Background noise poses a serious problem for both voice-based communication and automated services. Speech quality and intelligibility can be seriously hindered and automatic speech recognition systems are far less robust to noise than humans. Speech enhancement algorithms can ameliorate to some extent the aforementioned problems.

In this thesis, we are concerned with the development of speech enhancement algorithms whose aim is the improvement of the quality of noisy speech. As the notion of speech quality can be rather abstract and multidimensional [25], we focus the scope of our work into two main objectives: the first is the reduction of the level of the background noise, trying at the same time to avoid the harmful speech enhancement artifact known as musical noise [9], which consists of short tonal bursts that appear in random frequencies. The second objective is to preserve speech as accurately as possible, while minimising the distortions introduced by the processing.

The speech enhancement algorithms we propose in this thesis all fall in the category of single channel speech enhancement. We will not be concerned with dual or multi channel speech enhancement algorithms [61, 76]. A brief overview of single channel speech enhancement will be given in the next section, where we will also identify the specific genre of this family of algorithms that we will pursue. In §1.2 we will

outline the main developments of this thesis and highlight the novelty of this work. The structure of the thesis will be presented in §1.3 and in §1.4 we will detail the publications that have been derived from this work to date.

1.1 Single channel speech enhancement

Single channel speech enhancement algorithms assume the existence of a single sensor that captures the noisy speech. Therefore, algorithms of this type have to estimate the noise statistics and enhance the speech from a single recording. This is in contrast to dual channel speech enhancement for example, where the existence of a noise reference is assumed (e.g. [61]). The single channel speech enhancement literature has produced a large number of algorithms, which can be classified in the following categories:

- Bayesian estimators of the speech Short Time Fourier Transform (STFT) (e.g. Ephraim and Malah [31], Martin [72], Wolfe and Godsill [99])
- Spectral subtraction (e.g. Boll [13], Lim and Oppenheim [63])
- Speech enhancement based on Hidden Markov Models (e.g. Ephraim [30])
- Subspace methods (e.g. Ephraim and Van Trees [33], Rezayee and Gazor [85])
- Kalman filters (e.g. Paliwal and Basu [78], Gannot et al. [37])

The algorithms we develop in this thesis fall in the first category, that is, they are based on Bayesian estimators of the STFT. A typical algorithm of this type first transforms the noisy speech signal in the short time frequency domain by means of an STFT. An optimal clean speech estimator is then applied to the noisy speech STFT coefficients, assuming some distribution for the coefficients of speech and noise. Finally, an inverse STFT is applied in order to retrieve the enhanced signal in the time domain.

Algorithms based on Bayesian estimation of the STFT take advantage of the solid background of Bayesian theory, unlike the spectral subtraction algorithms for example, whose derivation has more empirical origins. The STFT is a computationally

cheap transformation, in contrast to the KLT transform that is typically applied in subspace methods. Finally, in an extensive comparison of single channel speech enhancement algorithms from various categories, which was presented by Hu and Loizou [52], the algorithms based on Bayesian estimation of the STFT were preferred by the majority of the subjects that participated in the listening tests.

1.2 Main developments

The work presented in this thesis can be divided in three major parts. In the first part (chapters 3 - 5), we develop a framework of Bayesian algorithms for speech enhancement, which consists of: i) generalisations of existing algorithms and ii) algorithms that are entirely novel. In the second part (chapter 6), we propose novel algorithms for the estimation of the noise power from a single channel recording of speech corrupted with noise. In the third and final part of this thesis (chapter 7), we employ tools from the theory of Markov Random Fields (MRF) for the development of speech enhancement algorithms. MRF's have found limited applications in speech processing so far and, to the best of our knowledge, this is the first time they are employed in enhancing speech corrupted with broadband noise.

The algorithms that comprise the proposed framework can be divided according to three of their features. The first is the clean speech STFT representation they estimate, which can be either the real (Re) and imaginary (Im) parts, or the amplitude. Secondly, they can be grouped according to the estimator they employ for the estimation of the STFT representation. The employed estimators are the Minimum Mean Square Error (MMSE) and the Maximum A Posteriori (MAP). A final possible grouping is according to the probability density function that is used for modelling the speech STFT coefficients (prior). The priors used with the algorithms that estimate the Re and Im parts are the 2 sided Chi and Gamma density functions. The priors used with the algorithms that estimate the amplitude are the 1 sided versions of the Chi and Gamma densities and the Lognormal density. A graphic representation of the algorithms that constitute the proposed framework, along with the 'code' names selected for each one, is shown in figure 1.1. The code names for the algorithms are based on the following format: the two first letters designate the estimator (i.e. MP for MAP and MS for MMSE). The next num-

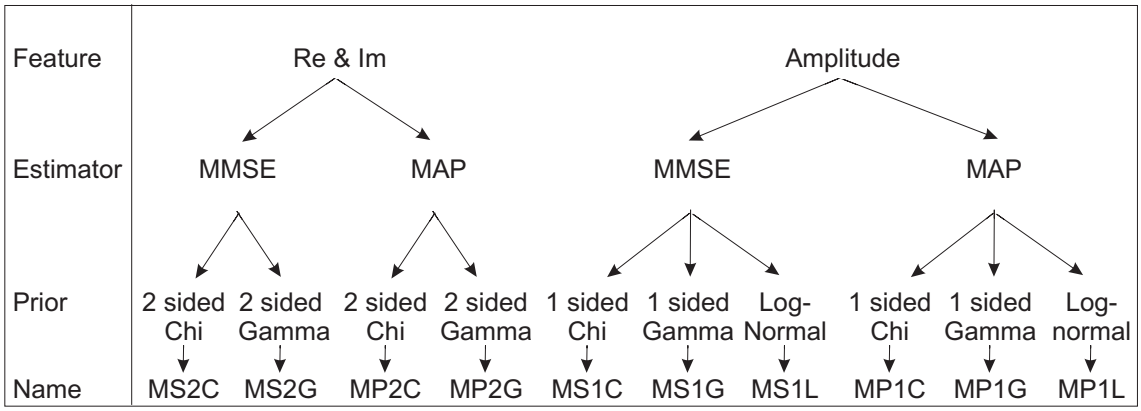


Figure 1.1: The proposed framework of Bayesian algorithms for speech enhancement.

ber determines whether the prior is 1 or 2 sided and also determines the estimated STFT representation as 2 sided priors are used for the estimation of the Re and Im parts and 1 sided priors are used for the estimation of the amplitude. Finally, the last letter denotes the name of the prior (i.e. C for Chi, G for Gamma and L for Lognormal).

A common characteristic of all the employed priors is a parameter that controls their shape, which we call a. Small values of a make the priors more leptokurtic (higher concentration around zero and longer tails), while large values of a result in more platykurtic priors. Apart from the above rather general observation however, we should mention that the same value of a has different effect in different priors (e.g. same a results in different values for the moments of the different priors). Nevertheless, the shape parameter a offers great flexibility in the shape the priors assume, and has an immense effect on the performance of the respective algorithms. The effect of a on the quality of the enhanced speech is a focal point of our research.

The proposed framework encapsulates several algorithms that exist in the literature, which are derived from the algorithms of the framework for particular values of a. These algorithms are: the Wiener filter [63], which is derived from the MS2C algorithm with $a = 1$, as the 2-sided Chi density with $a = 1$ is the Gaussian density. The Ephraim - Malah MMSE STSA algorithm [31] is derived from the MS1C with $a = 2$ as the 1 sided Chi density with $a = 2$ corresponds to the Rayleigh density used in [31]. Two of the algorithms proposed by Martin [72] are given by the MS2G algorithm with $a = 0.5$ and $a = 1$. A MAP algorithm proposed by Wolfe

and Godsill [99] is the MP1C with $a = 2$, while Lotter and Vary [66] proposed an algorithm which is the MP1G with $a = 2$. Finally, Dat et al. [24] proposed an instance of the MP1C and MP1G algorithms with $a = 1$ and $a = 1.5$ respectively. No instances of the MP2C, MP2G, MS1G, MS1L, MP1L have ever been proposed in the literature.

Apart from the introduction of algorithms that surpass the performance of the existing ones, the compilation of the above framework of algorithms has two additional benefits. Firstly, it provides the opportunity of directly comparing several popular speech enhancement algorithms that already exist in the literature. Secondly, and perhaps more importantly, due to the large number of algorithms that comprise the framework and the various groups they can be classified into, it offers an insight into the effect the different elements (estimator, prior etc.) have on the quality of the resulting speech, and yields interesting conclusions on their relative importance.

In the second part of this thesis, we present our work on the development of algorithms that estimate the power of time varying noises. The first part of this work, investigates the applicability of Gaussian Mixture Models in modelling the STFT coefficients of time varying noise. We show that the flexibility of these models allows an accurate modelling of the STFT coefficients of time varying noise, which motivates their employment in a speech enhancement scheme. In the second part, a noise estimation algorithm based on a single Gaussian distribution is developed, which exploits an observation that has received little attention in the literature. This observation regards the similarities between the distribution of the noisy speech spectral amplitude coefficients within a single frequency bin and the distribution of the respective coefficients of the corrupting noise. Taking advantage of the above similarities, we developed an algorithm that extracts from a window of past spectral amplitude samples of noisy speech those samples that are more likely to correspond to noise. The latter samples are then used to produce an noise power estimate. The extraction of the samples that belong to noise is based on matching the two first moments of the Rayleigh distribution.

Finally, in the last part of this thesis we investigate the applicability of MRF's to the problem of speech enhancement. MRF's have found extensive application in image processing problems, due to their ability to model interactions between

neighbouring pixels. Speech signals are known to have dependencies both in time and in frequency, which in the STFT domain manifest themselves as dependencies between neighbouring STFT samples. We therefore try to take advantage of the neighbour - modelling capabilities of the MRF's to develop speech enhancement algorithms that incorporate the time and frequency dependencies of speech signals.

1.3 Structure of the thesis

In chapter 2 we review the most prominent approaches for single channel speech enhancement. Naturally, we focus our attention on the methods based on Bayesian estimation of the STFT, making a distinction between the algorithms that are encapsulated in our framework and those which cannot be considered as its members, although their derivation is based on very similar principles. In chapter 2 we also present the basic concepts of Bayesian estimation, which form the stepping-stones to developments that follow.

In chapter 3 we formulate the problem of enhancing noisy speech as an estimation problem within a Bayesian context. After presenting analytically the proposed speech priors we derive the respective estimators for all the algorithms of the framework. At the end of the chapter we also attempt to verify the independence between the Re and Im parts and between the amplitude and phase of speech STFT coefficients, which is assumed throughout the chapter.

The proposed priors have two parameters: the shape parameter a and the scale parameter θ. Chapter 4 discusses various methods for their estimation. The discussed methods are grouped in two categories. The first is based on fitting the priors to a large number of clean speech data by means of minimising the Kullback-Leibler divergence. This method apart from providing a set of values that can be used for enhancing speech, can also give a measure for the appropriateness of the priors to model the speech data. The second group of methods estimate the values of the priors' parameters adaptively as the processing of noisy speech progresses. The adaptive method we employ for the estimation of the scale parameter is the Decision Directed (DD) method of Ephraim and Malah [31], while for the shape parameter a the method we propose is based on moment matching.

In chapter 5 we provide an extensive evaluation of the algorithms that comprise our framework. First, we investigate their performance as a function of the priors' shape parameter a and draw conclusions about its effect on the quality of speech. Optimal values for a are then sought by means of a formal subjective listening test. Finally, the adaptive scheme for the estimation of a is evaluated.

Our work on the development of noise estimation algorithms is presented in chapter 6. We begin by reviewing some of the most popular noise estimation methods, presenting them according to the principles on which they are based. The speech enhancement algorithm that employs the Gaussian Mixture Models of noise is then developed and evaluated. Finally, we describe the principles on which the Rayleigh moment matching noise estimation algorithm is based and its derivation is given in detail, followed by a comparison of the algorithm's performance with that of a state of the art noise estimation method.

Chapter 7 is a study on the applicability of MRF's to speech enhancement. We begin by laying down the theoretical background of the MRF's and then derive a MAP estimator of clean speech that is based on Gaussian MRF priors. We then introduce a novel type of MRF, which we term Chi MRF and employ it in the problem of speech enhancement. Finally, we discuss a limitation of using MRF's with fixed weights between the neighbours for speech enhancement and make an attempt to overcome them by introducing an adaptive scheme.

Finally, chapter 8 summarises the work presented in this thesis and draws the conclusions that have stemmed from this work. Additionally, directions into which this work could further expand are also given.

1.4 Novel contributions and publications

The main contributions of this work in the field of speech enhancement are the following:

- The generalisation of existing speech enhancement algorithms (see chapter 3: MS2C, MS2G, MS1C, MP1C and MP1G).

- The introduction of novel speech enhancement algorithms (see chapter 3: MP2C, MP2G, MS1G, MS1L and MP1L).
- The compilation of a framework of Bayesian algorithms for speech enhancement, which offers insight on the relative importance of the estimator, prior and estimated STFT feature.
- A noise estimation algorithm based on Gaussian Mixture Models.
- A noise estimation based on matching the moments of the Rayleigh distribution.
- The incorporation of MRF's for modelling the speech spectral amplitude.
- The introduction of Chi MRF's.
- The development of an adaptive scheme for the estimation of the MRF parameters that allows the restoration of the speech spectral components, while effectively suppressing the background noise.

The following is a list of publications that have arisen from this work to date.

- I. Andrianakis and P. R. White, *"Bayesian algorithms for speech enhancement"* ISVR Technical Report, No 305, Jan. 2006.
- I. Andrianakis and P. R. White, *"MMSE speech spectral amplitude estimators with Chi and Gamma speech priors"* in International Conference on Acoustics, Speech and Signal Processing (ICASSP-06), vol. 3, pp 1068-1071 May 2006.
- I. Andrianakis and P. R. White, *"Noise estimation based on matching the moments of the Rayleigh distribution for speech enhancement"* in Hellenic Institute of Acoustics 2006, Heraklion, Greece, Sep. 2006.
- I. Andrianakis and P. R. White, *"On the application of Markov Random Fields to speech enhancement"* in Proc. 7th IMA Int. Conf. Mathematics in Signal Processing, Cirencester, UK, Dec. 2006.

Chapter 2

Literature review and background material

In the past three decades numerous algorithms have been developed for the enhancement of noisy speech. The different approaches can be grouped according to the theory on which they are based into categories such as spectral subtraction algorithms, methods based on the Bayesian estimation of the STFT, signal subspace approaches, Hidden Markov Models (HMM), Kalman filtering etc.. In this chapter we present an overview of algorithms that belong in the different groups, while maintaining our focus on those which are based on the Bayesian estimation of the STFT, as these are the methods that are central to this thesis.

The algorithms that are based on the estimation of the STFT are reviewed in §2.1. We make a distinction between those that immediately fit into the proposed Bayesian framework of speech enhancement algorithms (§2.1.1) and those that although cannot be considered as its members, they can either be seen as its extensions or they are derived from a similar underlying theory. The latter group is presented in §2.1.2. In §2.2, the algorithms that are derived from alternative theoretical backgrounds (e.g. signal subspace, HMM's, Kalman filters) are presented. We provide a more detailed presentation of the algorithm(s) that triggered the research interest in the particular area, with the addition of some extensions that attempted to mitigate the shortcomings of the original methods.

Finally, in §2.3 we present the basic concepts of Bayesian estimation, which are

considered standard textbook material (e.g. [96]) and are fundamental in the development of the algorithms proposed in this thesis. Quantities such as the prior and posterior distributions will be defined and an analytical derivation of the MMSE and MAP estimators will be presented.

2.1 Methods based on Bayesian estimation of the STFT

A large number of speech enhancement methods that exist in the literature are based on Bayesian estimation of the clean speech STFT. These methods typically transform the noisy signal into the STFT domain and with the assumption of a statistical model produce an estimate of the clean speech STFT. The resulting estimate is then transformed back to the time domain in order to yield the enhanced speech signal.

A number of the methods that fall in the above category consist a subset of the algorithms from the Bayesian framework, which is proposed in this thesis. Additionally, there are a number of algorithms, which although cannot be directly incorporated into the above framework, are intimately linked with it. The two above categories of algorithms will be discussed in the next two sections.

2.1.1 Methods that belong in the proposed Bayesian framework

The framework we propose in this thesis consists of algorithms that i) estimate either the Re and Im parts or the amplitude of the speech STFT, ii) use the Chi, Gamma and Lognormal speech priors and iii) employ the MMSE or the MAP estimators. A number of algorithms which are contained in the proposed framework can be found in the literature. These are discussed in the following along with some motivation for their development.

One of the earliest algorithms that is a member of the above framework is the Wiener filter, which was first presented in the context of speech enhancement by Lim and Oppenheim [63]. The same algorithm was put in its Bayesian context by Martin [72], where it was explicitly stated that the Wiener filter is the MMSE estimator of the Re and Im parts of the speech STFT coefficients. The Re and Im parts of speech are assumed to follow a Gaussian distribution, which is an instance of the 2 sided

Chi density (eq. 3.7) when its shape parameter a takes the value 1. Capitalising on the importance of the short time speech amplitude relative to the short time phase in speech perception, Ephraim and Malah [31] developed the MMSE STFT amplitude estimator. The speech spectral amplitude was modelled with a Rayleigh distribution, which is an instance of the 1 sided Chi density (eq. 3.21) when its shape parameter a has the value 2.

Observing that the Re and Im parts of the speech STFT are better modelled by supergaussian densities, Martin [72] developed MMSE estimators of the Re and Im parts using instances of the 2 sided Gamma distributions (eq. 3.14) with $a = 1$ (Laplacian) and $a = 0.5$. Approximate MAP estimators of the speech spectral amplitude were then developed in [24, 66, 99]. Wolfe and Godsill [99] used the 1 sided Chi priors with $a = 2$ (Rayleigh), Lotter and Vary [66] used the 1 sided Gamma priors (eq. 3.28) with $a = 2$ and Dat et al. [24] used the 1 sided Gamma priors with $a = 1.5$ and the 1 sided Chi with $a = 1$.

2.1.2 Methods adjacent to the proposed framework

Apart from the algorithms mentioned in the previous section, there are also a number of algorithms that although do not immediately fit into the above framework, they do bear a large degree of similarity with its members. An example is the algorithm proposed by Porter and Boll [81], where the MMSE estimator of the speech spectral amplitude was developed under the same assumptions as in [31]. However, rather than assuming a closed form density function for the distribution of the amplitude coefficients, the authors proposed that the MMSE estimator is implemented empirically via the sample distribution of the clean speech signal. An obvious drawback of this method is that significant memory resources are required for the storage of the clean speech database. Ding et al. [28] developed the MMSE estimator of the squared speech spectral amplitude based on the assumption that the speech DFT coefficients are distributed according to a mixture of Gaussian distributions. Lotter and Vary [67] proposed a joint spectral amplitude and phase MAP estimator using 1 sided Gamma priors (eq. 3.28) and $a = 1.1$, while Gazor and Zhang [38] derived MMSE and MAP estimators for the Discrete Cosine Transform (DCT) representation of the speech signals, assuming that the speech coefficients follow a Laplacian

distribution and the coefficients of noise are Gaussian.

The following four studies introduced an estimator of the speech spectral amplitude based on the Gaussian assumption for the distribution of the speech and noise spectral coefficients. However, rather than using the amplitude mean square error (MSE) cost function, other cost functions were proposed. Ephraim and Malah [32] proposed the minimisation of the MSE of the logarithm of the spectral amplitude, while Cohen [18] combined the above estimator with the speech presence uncertainty method which was developed earlier by Ephraim and Malah [31]. You et al. [100] proposed the minimisation of the MSE of the spectral amplitude raised to an arbitrary power β and finally, Loizou [65] proposed a number of perceptually motivated cost functions. It is interesting to note, that the algorithm in [65] with the best overall performance is identical to our MMSE amplitude spectral estimator with the 1 sided Chi priors, despite the different motivation for their derivation.

McAulay and Malpass [73] adopted a somewhat different approach by modeling the speech spectral amplitude, not as a random variable, but as a deterministic complex variable with unknown amplitude and phase. Assuming then that the noise coefficients have a Gaussian distribution they derived a Maximum Likelihood estimator of the speech spectral amplitude. Hendriks et al. [47] used a model for the Re and Im parts of the speech STFT that consisted of a random plus a deterministic component. The random part of the model was used for noise like speech sounds such as the fricatives /s/, /f/, while the deterministic part was used for vowels. Estimators of the Re and Im parts of the DFT were then derived that involved a soft and a hard decision between the two parts of the model.

The speech enhancement method proposed by Tsoukalas et al. [94] utilised a psychoacoustic mechanism, known as noise masking. According to this, there exists a spectral amplitude threshold, called the Auditory Masking Threshold (AMT), below which all frequency components are masked in the presence of the masker signal (i.e. speech). The authors of [94] used the AMT in order to estimate the audible noise spectrum, which consists of those spectral components that are perceived as noise. The enhanced speech was then obtained with a parametric Wiener-type filter, whose parameters were estimated so that the audible noise spectrum is equal or less than zero. Extending the above work, Hansen et al. [45] proposed a statistical method for

the estimation of the AMT. Rather than using the heuristic iterative method of [94] for the estimation of the speech spectrum, which is required for the calculation of the AMT, an MMSE estimator of the speech power spectrum was used instead. A further extension to the above line of research was provided by You et al. [101], who employed their β power MMSE amplitude estimator [100] both for enhancing speech and for the estimation of the AMT. Additionally, the authors proposed a scheme for the on line adaptation of the value of β, which was based on the frame SNR and the estimated frame AMT, while an adaptation of a spectral flooring similar to that proposed by Virag [97] was also employed.

2.2 Alternative methods for speech enhancement

Despite the fact that the algorithms presented in the previous section include some of the most popular approaches for speech enhancement, they by no means exhaust the vast number of speech enhancement algorithms that exist in the scientific literature. In this section we summarise some of the most prominent alternative approaches.

2.2.1 Spectral Subtraction

An intuitive and simple in its implementation method for speech enhancement is the spectral subtraction. Its basis relies on the fact that if speech and noise are additive and uncorrelated, then the power spectral density of the noisy speech is equal to the sum of the power spectral densities of speech and noise [63]. If we denote by $X(k)$, $S(k)$ and $N(k)$ the amplitude of the DFT of a short segment of noisy speech, speech and noise respectively, then an estimate of the clean speech DFT amplitude $\hat{S}(k)$ can be obtained as

$$\hat{S}^{\gamma_{\mathrm{SS}}}(k) = \max(\alpha_{\mathrm{SS}} X^{\gamma_{\mathrm{SS}}}(k) - \beta_{\mathrm{SS}} \hat{N}^{\gamma_{\mathrm{SS}}}(k), \delta_{\mathrm{SS}}\, \hat{N}^{\gamma_{\mathrm{SS}}}(k)) \tag{2.1}$$

where k is the frequency bin index and $\hat{N}(k)$ is an estimate of the noise spectrum. α_{SS}, β_{SS}, γ_{SS} and δ_{SS} are all positive parameters. In particular, β_{SS} is known as the oversubtraction factor, which determines the amount of the subtracted noise. The exponent γ_{SS} controls the aggressiveness of the algorithm, resulting in lower levels

of residual noise but higher speech distortion as it approaches zero. Finally, δ_{SS} controls the noise floor, which is a minimum value for the spectral estimates and can aid the suppression of musical noise.

The method proposed by Boll [13] used $\alpha_{\text{SS}} = \beta_{\text{SS}} = \gamma_{\text{SS}} = 1$ and $\delta_{\text{SS}} = 0$, a method known as amplitude spectral subtraction. The method proposed by Lim and Oppenheim [63] used the same parameter values as above except for $\gamma_{\text{SS}} = 2$, which is known as power spectral subtraction. Although these methods result in sensible estimates of the clean speech, they both suffer from high levels of musical residual noise. To alleviate this problem, Boll [13] proposed to replace the current spectral value estimate in a time frame with the minimum of the adjacent frames, exploiting in this way the random nature of musical noise. Berouti et al. [9] proposed the use of $\beta_{\text{SS}} > 1$ and $\delta_{\text{SS}} = 0.01$, in order to reduce the amount of perceived musical noise. The authors of [9] also experimented with arbitrary powers of γ_{SS}, concluding that the optimum results were obtained for $\gamma_{\text{SS}} = 2$. Scalart and Filho [88] proposed to incorporate the DD method for the estimation of the a priori SNR [31] in the power spectral subtraction method. This was motivated by the success of the DD method in suppressing musical noise as reported in [16]. Sim et al. [90] proposed the use of estimates of α_{SS} and β_{SS} that minimised the mean square error between $S(k)$ and $\hat{S}(k)$. Finally, Virag [97] used the perceptual model employed by Tsoukalas et al. [94] for the calculation of the AMT, which was then used in the adaptive estimation of β_{SS} and δ_{SS}. The latter parameters were adapted in such a way that less suppression was applied when the value of the AMT was high, in order to minimise the speech distortion, taking also into account that the noise should be masked anyway by the speech signal for high values of the AMT.

2.2.2 Hidden Markov Models

A speech enhancement method that is based on Hidden Markov Models (HMM) was proposed by Ephraim [30] and references therein. The proposed method is based on the Bayesian framework that was presented in §2.1, with the difference that the speech and noise signals were modelled with HMM's instead of simple density functions. The speech and noise signals were modelled with first order HMM's with Gaussian state dependent density functions, each of which was assumed to be an

AR process. Given the speech and noise HMM models, MMSE and MAP estimators were then derived, in a similar fashion as in §2.3. The resulting estimators comprised of $\mathcal{M} \times \tilde{\mathcal{M}}$ Wiener filters, where $\mathcal{M}$ is the number of HMM states for the speech signal and $\tilde{\mathcal{M}}$ the states for the noise signal. The clean speech estimate was obtained with different combinations of the $\mathcal{M} \times \tilde{\mathcal{M}}$ Wiener filters, which was determined from the employed estimator (MMSE or MAP).

The need for the calculation of $\mathcal{M} \times \tilde{\mathcal{M}}$ Wiener filters imposes an increased computational load to the HMM based methods. Additionally, their complexity can be somewhat increased as the parameters of the HMM models need to be calculated from training on speech and noise databases, while their performance depends on the match between the training and test data [21].

2.2.3 Subspace Methods

The subspace methods are based on the decomposition of the noisy speech signal into two subspaces: the speech plus noise and the noise only subspace. Once the decomposition is achieved, the noise subspace is discarded, while the clean speech is estimated from the remaining speech plus noise subspace.

The mixing model of speech and noise is given by

$$\mathbf{x} = \mathbf{s} + \mathbf{n} \tag{2.2}$$

where $\mathbf{x}$, $\mathbf{s}$ and $\mathbf{n}$ are vectors of noisy speech, speech and noise respectively that contain K_x samples each. The model that is assumed for the speech signal is

$$\mathbf{s} = \mathbf{W}\mathbf{y} \tag{2.3}$$

where $\mathbf{y}$ is a K_s dimensional vector of zero mean random variables. The matrix $\mathbf{W}$ consists of K_s basis vectors, whose dimension is K_x. The fundamental assumption of the subspace family of methods is that a K_x dimensional speech vector can be represented as a linear combination of $K_s < K_x$ basis vectors. Under this assumption, the vector $\mathbf{s}$ lies in a subspace $\Re^{K_s}$ of the Euclidean space $\Re^{K_x}$, which is spanned by the columns of the matrix $\mathbf{W}$ and is called speech or speech plus noise subspace.

The covariance matrix of $\mathbf{s}$ is

$$\Sigma_s \equiv \mathrm{E}\left[\mathbf{s}\mathbf{s}^{\#}\right] = \mathbf{W}\Sigma_y\mathbf{W}^{\#} \tag{2.4}$$

where Σ_y is the covariance of $\mathbf{y}$ and $(.)^{\#}$ denotes conjugate transpose. The rank of the Σ_s matrix is K_s, which implies that it has K_s positive, and $K_x - K_s$ zero, eigenvalues.

For the K_x dimensional noise vector $\mathbf{n}$ on the other hand, it is assumed that its covariance matrix $\Sigma_n \equiv \mathrm{E}\left[\mathbf{n}\mathbf{n}^{\#}\right]$ has a rank of K_x with K_x positive eigenvalues. In other words, the noise vectors fill the entire Euclidean space $\Re^{K_x}$, which consists of the signal subspace $\Re^{K_s}$ and its compliment $\Re^{K_x-K_s}$. The latter is called noise subspace.

The method proposed by Ephraim and Van Trees [33] achieved the decomposition into speech and noise subspaces by means of a Karhunen-Loeve Transform (KLT). Two estimators of the speech signal from the speech subspace were then developed: the first minimised the speech distortion, while keeping the noise energy within each frame below a certain threshold (time domain constrained estimator), and the second estimator minimised the speech distortion, while keeping below a threshold the energy in each spectral component (spectral domain constrained estimator). The resulting estimators were closely related to the Wiener filter.

A drawback of the above approaches is that they are designed for white noise, while colored noise can be handled only with a prewhitening step. A number of extensions to the above methods have been proposed since, that can explicitly handle colored noise. Mittal and Phamdo [74] proposed to classify the speech frames as containing mainly speech or noise and apply the KLT to the dominant process in each frame. Rezayee and Gazor [85] proposed the use of a diagonal matrix (as opposed to the identity matrix of [33]) for the approximation of the colored noise spectrum. Hu and Loizou [49] proposed the simultaneous diagonalisation of the speech and noise covariance matrix with a non orthogonal transformation. The simultaneous diagonalisation of the two covariance matrices with an orthogonal transformation was achieved by Lev Ari and Ephraim [62].

Jabloun and Champagne [54] enhanced the spectral domain constrained estimator

of [33] with the psychoacoustic models proposed in [94,97]. The AMT's were first estimated by obtaining an estimate of the clean speech covariance matrix $\Sigma_s = \Sigma_x - \Sigma_n$, where Σ_x is the covariance matrix of the noisy signal and applying an eigendomain to frequency transformation. Subsequently, using a frequency to eigendomain transformation, a set of eigenvalues that contained the perceptual information of the psychoacoustic model were calculated, which were then used in the estimation of the clean signal. An alternative method for the incorporation of a psychoacoustic model with the subspace algorithms was also proposed by Hu and Loizou [50]. Rather than using the AMT's, an estimate of the speech spectrum was obtained with an LPC polynomial, whose inverse spectrum was used to perceptually weight the time domain error signal. A clean speech estimator similar to the spectral domain constrained estimator of [33] was then developed, which minimised the perceptually weighted error criterion rather than the mean squared error.

Apart from the above KLT-based methods, Dendrinos et al. [27] and Jensen et al. [55] proposed methods which are based on the Singular Value Decomposition (SVD). The method in [27] is an SVD-based method similar to the first estimator of [33], while the method in [55] is a colored noise extension to the method in [27].

A drawback of the subspace methods is that they are computationally more demanding than the STFT estimation methods (§2.1). The load is imposed by the relatively expensive computation of the KLT or SVD transforms. Additionally, in the extensive subjective comparison of different classes of algorithms, which was presented in [52], the subspace algorithms obtained lower scores compared to the algorithms based on Bayesian estimation of the STFT.

2.2.4 Kalman Filters

Another family of speech enhancement algorithms is based on Kalman filters. In this family of algorithms the time domain speech samples $s(i)$ are typically modelled with an AR process of the form

$$s(i) = \sum_{m=1}^{N_{\mathrm{AR}}} a_r(m,i)s(i-m) + e_r(i) \tag{2.5}$$

where $a_r(m,i)$ are the time varying AR coefficients, N_{AR} is their number and $e_r(i)$ is the driving noise sequence. Based on the above speech model and the linear mixing model

$$x(i) = s(i) + n(i) \tag{2.6}$$

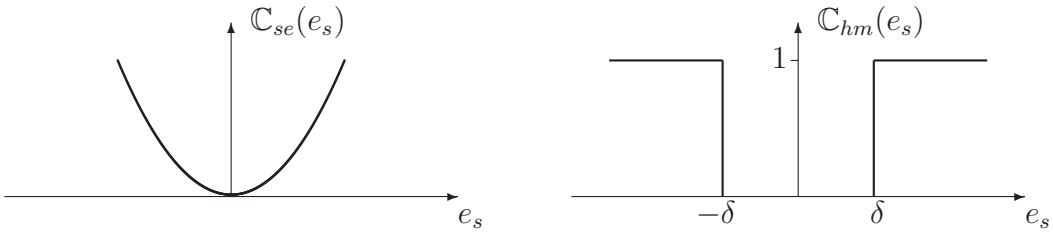
where $x(i)$, $s(i)$ and $n(i)$ are the noisy speech, speech and noise signals, the equations of the standard Kalman filtering for the estimation of clean speech were proposed by Paliwal and Basu [78], under the assumption that both the driving noise sequence $e_r(i)$ and the noise signal $n(i)$ are white and zero mean.

An extension to colored noise was proposed by Gibson et al. [40], by incorporating an AR model for noise in the state equations of the Kalman filter. Gannot et al. [37] enhanced the previous combined speech and noise model with the addition of the Estimation Maximisation (EM) algorithm for the estimation of the speech and noise model parameters. Incorporating a psychoacoustical model, Ma et al. [68] derived a Kalman filter under the constraint that the estimation error is smaller than a masking threshold, while both simultaneous frequency masking and time domain masking were taken into account.

Finally, by taking a slightly different approach, Zavarehei et al. [102] proposed the use of Kalman filters for estimating the Re and Im parts of the speech STFT. Estimators where the noise was modelled as either an uncorrelated or an AR process were then developed, and results comparable with those of well known Bayesian STFT estimators were achieved.

2.3 Bayesian estimation

In this section we discuss the theoretical background of Bayesian estimation, which will be central in the development of the proposed speech enhancement algorithms. We will introduce quantities such as the prior and posterior distributions and will derive the MMSE and MAP estimators, which will be extensively employed in the following chapters. The section will close with an example of estimating a random variable buried in noise, in an attempt to further clarify the various concepts of Bayesian estimation and the procedure itself. A more comprehensive treatment of the material presented in this section can be found in [96].



Square error Hit-or-miss

Figure 2.1: Typical cost functions.

A central concept in Bayesian estimation is the *cost* function $\mathbb{C}\left(s, \hat{s}(x)\right)$, where s is the random variable (r.v.) we are trying to estimate, x is the observed r.v. and $\hat{s}(x)$ is an estimate of s once x is observed. The cost function defines the cost of observing x and saying that the estimate for s is $\hat{s}(x)$. It is often possible to express the cost as a function of a single variable $e_s(x)$, which is called the error and is defined as

$$e_s(x) = \hat{s}(x) - s \tag{2.7}$$

Typical cost functions include the square error (eq. 2.8) and the 'hit-or-miss' cost function (eq. 2.9), which assigns a uniform cost for absolute error values above a threshold δ. The above cost functions are illustrated in figure 2.1, while their analytical expressions are given below.

$$\mathbb{C}_{se}(e_s) = e_s^2 \tag{2.8}$$

$$\mathbb{C}_{hm}(e_s) = \begin{cases} 0 & \text{if } |e_s| < \delta \\ 1 & \text{if } |e_s| > \delta \end{cases} \tag{2.9}$$

Once a cost function is chosen, the objective is to minimise its expected value. The expectation (average) is with respect to all the possible values of the r.v.'s s and x and is often referred to as the *risk* $\mathbb{R}$, which is defined in eq. 2.10. $p(s, x)$ is the joint probability density function (joint pdf) of s and x.

$$\mathbb{R} \equiv \mathrm{E}[\mathbb{C}\left(e_s(x)\right)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{C}\left(e_s(x)\right) p(s, x)\; ds\, dx \tag{2.10}$$

Minimisation of the risk for different cost functions leads to different estimators. The estimators that are derived when the square error and hit-or-miss cost functions are

used are the Minimum Mean Square Error (MMSE) and Maximum A Posteriori (MAP) estimators respectively. These estimators are the principal ones used in practice and will be examined in the following sections.

2.3.1 Minimum Mean Square Error estimator

The MMSE estimator is obtained by minimising the risk function (eq. 2.10) with respect to $\hat{s}(x)$, using the square error cost function (eq. 2.8). The risk function can be written as

$$\mathbb{R} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (s - \hat{s}(x))^2 \, p(s, x) \; ds \, dx \tag{2.11}$$

Application of Bayes' theorem transforms the above equation to

$$\mathbb{R} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (s - \hat{s}(x))^2 \, p(s|x) \; ds \; p(x) \; dx \tag{2.12}$$

As $p(x)$ and the inner integral are non-negative, minimising the latter with respect to $\hat{s}$ also minimises the risk. Differentiation of the inner integral w.r.t. $\hat{s}$ yields

$$\frac{d}{d\hat{s}} \left[\int_{-\infty}^{\infty} (s - \hat{s}(x))^2 \, p(s|x) \; ds \right] = -2 \int_{-\infty}^{\infty} (s - \hat{s}(x)) \, p(s|x) \; ds \tag{2.13}$$

Setting eq. 2.13 to zero and considering that the integral of $p(s|x)$ from $-\infty$ to ∞ is 1, we see that the estimate that minimizes the mean square error is

$$\hat{s}(x) = \int_{-\infty}^{\infty} s \; p(s|x) \; ds = \mathrm{E}[s|x] \tag{2.14}$$

where $\mathrm{E}[s|x]$ is the conditional statistical expectation of s given x. It is interesting to note that the MMSE estimate is always the mean of the a posteriori density $p(s|x)$ (see figure 2.3). Further application of the Bayes theorem on eq. 2.14 can yield the expression in eq. 2.15, where the MMSE estimate is expressed in terms of the likelihood $p(x|s)$ and the prior $p(s)$ densities.

$$\hat{s}(x) = \frac{\displaystyle\int_{-\infty}^{\infty} s \; p(s, x) \; ds}{p(x)} = \frac{\displaystyle\int_{-\infty}^{\infty} s \; p(s, x) \; ds}{\displaystyle\int_{-\infty}^{\infty} p(s, x) \; ds} = \frac{\displaystyle\int_{-\infty}^{\infty} s \; p(x|s) p(s) \; ds}{\displaystyle\int_{-\infty}^{\infty} p(x|s) p(s) \; ds} \tag{2.15}$$

2.3.2 Maximum A Posteriori estimator

The maximum a posteriori estimator can be found by substituting the hit-or-miss cost function (eq. 2.9) in the expression for the risk (eq. 2.10), which then reads

$$\mathbb{R} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{C}_{hm}(e_s(x)) p(s,x)\ ds\, dx \tag{2.16}$$

Applying Bayes' rule and following the same argument as in eqs. 2.12 and 2.13 we see that for the minimisation of the risk it suffices to minimise

$$\mathbb{R}' = \int_{-\infty}^{\infty} \mathbb{C}_{hm}(e_s(x)) p(s|x)\ ds \tag{2.17}$$

Considering that the cost function $\mathbb{C}_{hm}(e_s(x))$ is 1 only for $e_s(x) > |\delta|$ or equivalently for $s > \hat{s}(x) + \delta$ and $s < \hat{s}(x) - \delta$ while it is zero everywhere else, eq. 2.17 can be written as

$$\mathbb{R}' = \int_{-\infty}^{\hat{s}(x)-\delta} p(s|x)\ ds + \int_{\hat{s}(x)+\delta}^{\infty} p(s|x)\ ds \tag{2.18}$$

or as

$$\mathbb{R}' = 1 - \int_{\hat{s}(x)-\delta}^{\hat{s}(x)+\delta} p(s|x)\ ds \tag{2.19}$$

if we recall that $\int_{-\infty}^{\infty} p(s|x)\ ds = 1$. As δ approaches zero, the value of $\hat{s}(x)$ that minimises $\mathbb{R}$ is the value of s for which $p(s|x)$ has its maximum. In other words, the risk is minimized for the hit-or-miss cost function when the estimate is the maximum (*mode*) of the posterior density function (see figure 2.3). Analytically, and with the application of Bayes' rule, this estimator can be written as

$$\hat{s}(x) = \arg\max_s p(s|x) = \arg\max_s \frac{p(x|s)p(s)}{p(x)} \tag{2.20}$$

Finally, by observing that $p(x)$ in the above equation does not depend on s the estimator takes the form

$$\hat{s}(x) = \arg\max_s p(x|s)p(s) \tag{2.21}$$

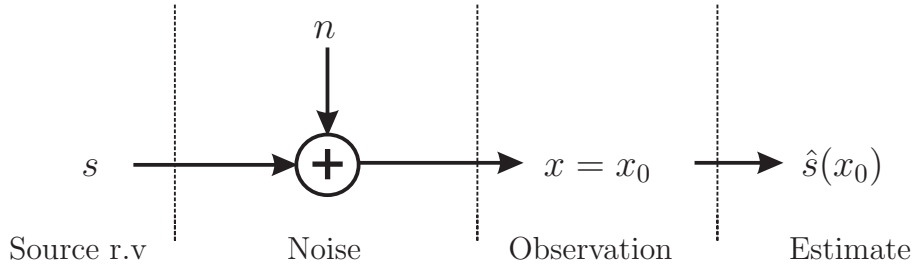


Figure 2.2: The r.v. s is corrupted additively by random noise noise n. By observing only the r.v. x, which takes on the value x_0, we seek to produce an estimate $\hat{s}(x_0)$ for the r.v. s.

2.3.3 An estimation example

In this section we present a simple estimation example for the clarification of the concepts introduced previously. Suppose that we have a random variable s which is corrupted with additive and independent noise n according to figure 2.2. We observe only their sum x, which takes the value x_0 and we seek an estimate $\hat{s}(x_0)$ of the r.v. s, which is a function of the observation x.

In order to produce an estimate according to the theory described in the previous sections we first need to define the prior distribution and the likelihood, which are denoted by $p(s)$ and $p(x|s)$ respectively. The prior distribution can be determined either by observing several realisations of s or by the possession of some knowledge about the generating process. For the purposes of this example let us assume that s follows a Laplacian distribution $p(s) = \frac{1}{2\theta} \exp\left[-\frac{|s|}{\theta}\right]$ as shown in figure 2.3(a). Suppose also that the noise follows a Gaussian distribution $p(n) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{n^2}{2\sigma^2}\right]$. The likelihood $p(x|s)$ can then be derived as follows: the joint density $p_{x,s}(x, s)$ can be obtained from the joint density $p_{n,s}(n, s)$ and a bivariate transformation $x = s+n$ and $s = s$ as ([79] p.201)[1]

$$p_{x,s}(x, s) = p_{n,s}(x - s, s) \tag{2.22}$$

The assumption of independence between s and n allows us to factorise $p_{n,s}(n, s)$,

[1]Subscripts have been introduced in the pdf's (i.e. $p_x(x)$) to maintain notational clarity. See also the last paragraph of this section for an explanation on the notation of the probability density functions.

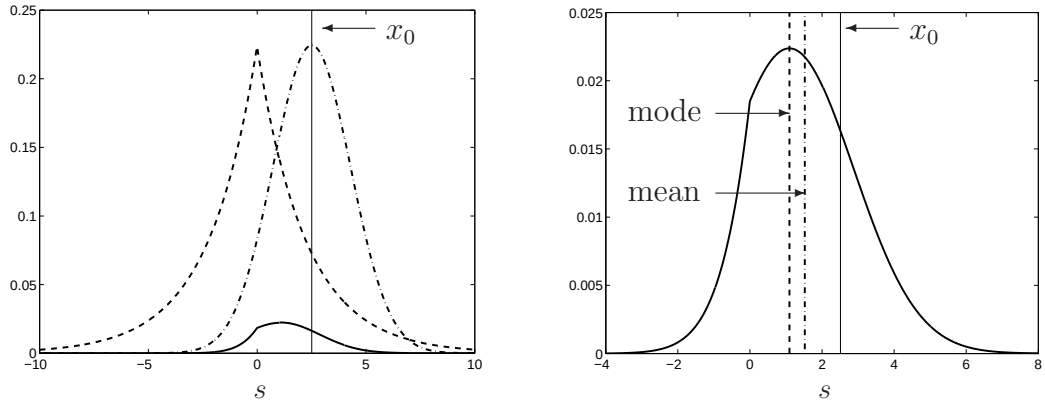


Figure 2.3: (a) Prior (dash), likelihood (dash dot) and posterior (continuous) densities. (b) The posterior density with the MAP (mode) and MMSE (mean) estimates. In both figures x_0 denotes the observation.

therefore eq. 2.22 can be written as

$$p_{x,s}(x,s) = p_n(x-s)p_s(s) \tag{2.23}$$

Application of Bayes' theorem in $p_{x,s}(x,s)$ yields

$$p_{x|s}(x|s)p_s(s) = p_n(x-s)p_s(s) \tag{2.24}$$

and finally

$$p_{x|s}(x|s) = p_n(x-s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(s-x)^2}{2\sigma^2}\right] \tag{2.25}$$

Hence, the likelihood is the distribution of the noise (Gaussian) centered at the value of the observation $x = x_0$. This distribution is also shown in figure 2.3(a).

In the above, the prior density encapsulates any prior knowledge we might have about the r.v. we are trying to estimate, while the likelihood represents the evidence provided by the data x. The product of the above two densities is related to the posterior density $p(s|x)$ according to Bayes' theorem through the equation

$$p(s|x) = \frac{p(s)p(x|s)}{p(x)} \tag{2.26}$$

where, if we consider s as the free r.v., then $p(x)$ is a mere normalising factor. The posterior distribution $p(s|x)$ is shown in figure 2.3(b). As already shown in §2.3.1

and §2.3.2, the MMSE and the MAP estimates are the mean and the mode of the posterior density respectively. The two different estimates are also shown in figure 2.3(b).

Before we close this section we make a comment on the notation of the probability density functions. The formal notation of a pdf requires a subscript and an argument i.e. $p_x(x_0)$. The subscript denotes the random variable the function refers to, while the argument is the independent variable of the function, which can be a mere number (i.e. $x_0 = 5$). For example, $p_x(x_0)$ denotes the probability density of the r.v. x at $x = x_0$. However, when there is no fear of ambiguity the subscript is dropped and the argument defines both the independent variable of the function and the random variable.

Chapter 3

A framework of Bayesian estimators of the speech STFT

The estimators presented in this chapter are used for the estimation of the STFT of the clean speech, when only the noisy speech STFT is observed. The proposed estimators can be categorised according to the STFT feature they estimate, the cost function they employ, and finally according to the speech prior density they assume. The STFT features considered here are the Re and Im parts or the amplitude of the STFT and the cost functions used are the squared error and the 'hit or miss' (see figure 2.1), which lead to the MMSE and MAP estimators correspondingly. The estimators of the Re and Im parts of the STFT use the 2 sided Chi and Gamma prior densities, while the amplitude estimators use the 1 sided versions of the above densities and additionally use the Lognormal priors. The assemblage of the above framework of estimators allows us to obtain an insight on the effect of the different components of an estimator to the quality of the enhanced speech, while it also encapsulates several successful speech enhancement algorithms, which can be found in the literature, as discussed in §2.1.1. After the presentation of each of the algorithms, their instances that already exist in the literature will be detailed.

The formulation of speech enhancement as an estimation problem is given in §3.1, while in the two next sections the estimators of the Re and Im parts and the amplitude estimators of the clean speech STFT are presented correspondingly. The development of the above two groups of estimators assumes that either the Re and

Im parts or the amplitude and the phase of the clean speech STFT are independent. The above assumptions cannot be valid simultaneously for distributions other than the complex Gaussian. The validity of the above assumptions for speech data is discussed in §3.4.

3.1 Problem formulation

Let us denote by $s(i)$ and $n(i)$ the sampled speech and noise signals, which are assumed to be independent and zero mean. The noisy speech signal $x(i)$ is modelled as the sum of $s(i)$ and $n(i)$. Although we acknowledge that real life noisy speech signals might be generated by a process more complex than the mere addition of the noise and speech signals (e.g. the Lombard effect), for the purposes of this thesis we assume that the electrical instantaneous mixing suffices. The transformation of $x(i)$ to the STFT domain is achieved by windowing the first K samples with a tapered window $h(i)$ of length K and applying an K point DFT to the windowed data. The window is then shifted by J samples and the procedure is repeated for the remainder of the signal. The STFT transformation can be written as

$$\mathbf{X}(k,l) = \sum_{m=0}^{K-1} x(Jl+m)\ h(m)\ e^{-\mathbf{i}2\pi\frac{mk}{K}} \tag{3.1}$$

where k is referred to as the frequency bin index and l as the time frame index.

According to the linearity property of the Fourier transform, the relationship between the STFT's of $x(i)$, $s(i)$ and $n(i)$ is

$$\mathbf{X}(k,l) = \mathbf{S}(k,l) + \mathbf{N}(k,l) \tag{3.2}$$

The task of speech enhancement algorithms is to produce an estimate of $\mathbf{S}(k,l)$ when only $\mathbf{X}(k,l)$ is observed. Half of the algorithms we present here estimate the Re and Im parts of $\mathbf{S}(k,l)$. In the following we will refer to these algorithms as the 'DFT algorithms' and they will be presented in the following section. The other half of the algorithms estimate the amplitude of the clean speech STFT, which is then combined with the phase of noisy speech to produce the enhanced speech signal. We will collectively refer to the latter group as the 'Amplitude algorithms' and they

will be introduced in §3.3. The amplitude algorithms that use the Chi and Gamma priors have been published in [7]. The amplitude and DFT algorithms that use the 1 and 2 sided Chi and Gamma priors respectively have been published in [4].

3.2 DFT algorithms

The algorithms we present in this section estimate the Re and Im parts of the clean speech STFT. The assumption of their independence allows their separate estimation, thus dividing the problem into two disjoint parts. To simplify the notation, X, S, and N will denote the real part of an STFT sample of the noisy speech, clean speech and noise respectively. For the three quantities it will also hold that $X = S + N$ as a result of eq. 3.2. In the following, we will derive the estimators for the Re parts of the involved STFT quantities, while the derivations for the Im parts are identical.

The estimation problem can be formulated as follows: we observe a sample of X and we want to estimate S given the noise and speech statistics. The derivation of the MMSE and MAP estimators requires the calculation of the posterior probability density function $p(S|X)$. According to Bayes' theorem, the posterior density can be written as

$$p(S|X) = \frac{p(X|S)p(S)}{p(X)} \tag{3.3}$$

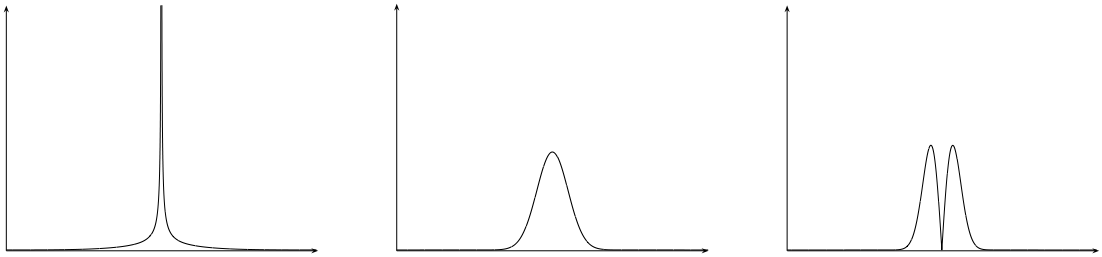
According to eq. 2.25 the likelihood $p(X|S)$ is given by

$$p(X|S) = p_N(X - S) \tag{3.4}$$

where p_N is the pdf of N. Assuming that N is a zero mean Gaussian r.v. with variance σ_N^2, the likelihood $p(X|S)$ can be written as

$$p(X|S) = \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left[-\frac{(X-S)^2}{2\sigma_N^2}\right] \tag{3.5}$$

The prior $p(S)$ is a density function that reflects our knowledge about the distribution of S. We will see in the following that the form of the prior strongly affects the performance of the resulting algorithm. The prior densities considered here are the



(a) $a = 0.1$ (b) $a = 1$ (c) $a = 2$

Figure 3.1: 2 sided Chi pdf's for different values of a.

2 sided Chi and Gamma pdf's that will be presented shortly.

The probability density of the data $p(X)$ in eq. 3.3 is a normalising factor that does not depend on S and ensures that the integral of the posterior density, with respect to S, is equal to 1. The density $p(X)$ can be calculated according to Bayes' rule as

$$p(X) = \int_{-\infty}^{\infty} p(X|S)p(S)\ dS \tag{3.6}$$

Note the similarity of the numerator and the denominator of eq. 3.3 if $p(X)$ is replaced from eq. 3.6.

3.2.1 2 sided Chi speech priors

The 2 sided Chi pdf is given by

$$p_{2C}(S) = \frac{1}{\theta^{a/2}\Gamma(a/2)}\ |S|^{a-1}\ \exp\left[-\frac{S^2}{\theta}\right] \tag{3.7}$$

where $\Gamma(.)$ is the gamma function. This is the 2 sided version of the Chi density with a degrees of freedom and scale parameter $\sqrt{\theta/2}$ [56]. Special cases of this distribution occur when $a = 1$ (Gaussian) and when $a = 2$ (2 sided Rayleigh). Figure 3.1 shows some instances of the 2 sided Chi pdf for some characteristic values of a.

3.2.1.1 MMSE estimator (MS 2C algorithm)

As shown in §2.3 the MMSE estimator is the mean of the posterior density. Therefore, the MMSE estimator of S will be

$$\hat{S} = \mathrm{E}\left[S|X\right] = \frac{\int_{-\infty}^{\infty} S\ p(X|S)\ p(S)\ dS}{\int_{-\infty}^{\infty} p(X|S)\ p(S)\ dS} \tag{3.8}$$

where $p(S)$ and $p(X|S)$ are given by eqs. 3.7 and 3.5 respectively. Calculation of the integrals in 3.8 yields (see appendix A.2)

$$\hat{S} = a\sigma_N^2\zeta\ \frac{D_{-a-1}(-\zeta X) - D_{-a-1}(\zeta X)}{D_{-a}(-\zeta X) + D_{-a}(\zeta X)} \qquad \text{where} \quad \zeta = \sqrt{\frac{\theta/\sigma_N^2}{\theta + 2\sigma_N^2}} \tag{3.9}$$

where $D_{\cdot}(.)$ is the Parabolic Cylinder Function (eq. 9.240, [42]). Its calculation is performed with the routine mpbdv.m found in [8]. For $|\zeta X| > 40$, where numerical problems typically occur in the calculation of $D_{\cdot}(.)$, the asymptotic expressions 9.246.1-3 found in [42] are used, producing numerically stable results for all input ranges.

Since their introduction in [31], the a priori SNR ξ and the a posteriori SNR γ have become an integral part of the speech enhancement literature. It is very often that speech enhancement estimators are expressed in the from of a gain function, whose arguments are the above two quantities. We will also follow the same practice here, for all the estimators that can be derived in a closed form. The definition of the a priori SNR is $\xi = \mathrm{E}[|\mathbf{S}|^2]/\mathrm{E}[|\mathbf{N}|^2]$ and the formula that relates ξ to the scale parameter θ of the 2 sided Chi density is $\xi = \theta a/2\sigma_N^2$ [1]. The definition of the a posteriori SNR in [31] was $\gamma = |\mathbf{X}|^2/\mathrm{E}[|\mathbf{N}|^2]$. This is a definition suitable for the estimators of the STFT amplitude, as it involves the term $|\mathbf{X}|^2$. For the DFT estimators we propose an alternative definition, which is $\gamma_2 = X^2/\mathrm{E}[N^2]$ or equivalently $\gamma_2 = X^2/\sigma_N^2$. Substituting the expressions for ξ and γ_2 in eq. 3.9 and

[1] The rationale behind the connection between the scale parameter θ and the a priori SNR ξ is given at the beginning of chapter 4, while the expressions that relate θ and ξ for all the considered priors are derived in §4.3.2

denoting by sgn(.) the signum function we obtain

$$\hat{S} = X\left[\frac{a\eta}{\gamma_2}\frac{D_{-a-1}(-\eta) - D_{-a-1}(\eta)}{D_{-a}(-\eta) + D_{-a}(\eta)}\right] \qquad \text{where} \quad \eta = \text{sgn}(X)\sqrt{\frac{\xi\gamma_2}{\xi + a}} \tag{3.10}$$

For $a = 1$ the MS2C algorithm is equivalent to the Wiener filter [63, 72], as we will discuss at the end of the following section.

3.2.1.2 MAP estimator (MP 2C algorithm)

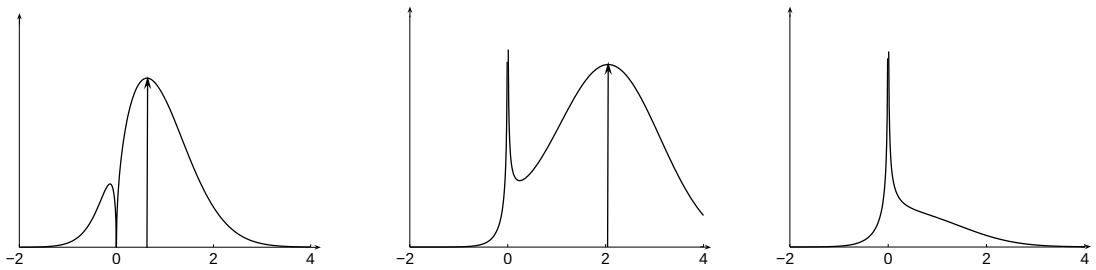
The MAP estimator $\hat{S}$ is the value of S for which the posterior density has its maximum. The probability of the data $p(X)$ is not a function of S so it suffices to find the maximum of $p(X|S)p(S)$, which are respectively defined by eqs. 3.5 and 3.7. The algebraic manipulations are substantially simplified if $\ln(p(X|S)p(S))$ is maximised. The resulting estimator is given by (see appendix A.3)

$$\hat{S} = \zeta\frac{X}{2} + \text{sgn}(X)\left[\left(\zeta\frac{X}{2}\right)^2 + (a-1)\,\sigma_N^2\,\zeta\right]^{1/2} \qquad \text{where} \quad \zeta = \frac{\theta}{\theta + 2\sigma_N^2} \tag{3.11}$$

It is also possible to express the above estimator as a gain for the noisy coefficients, which is a function of the a priori and a posteriori SNR, as they were defined in §3.2.1.1. The resulting expression is

$$\hat{S} = X\left[\frac{\eta}{2} + \left[\left(\frac{\eta}{2}\right)^2 + (a-1)\frac{\eta}{\gamma_2}\right]^{1/2}\right] \qquad \text{where} \quad \eta = \frac{\xi}{\xi + a} \tag{3.12}$$

For $a < 1$ the 2 sided Chi density function (eq. 3.7) has a singularity at zero, which the posterior density, given in eq. 3.3, inherits. The existence of the singularity in the posterior density implies that the global maximum is at zero. The use of zero as an estimate however, does not result in a useful algorithm. The strategy we follow in this case is to take the local maximum provided by eq. 3.11 when it exists and when it does not (or when the argument of the square root is negative) we suppress X by a fixed amount (i.e. 50 dB). Figure 3.2 shows three instances of the posterior density $p(S|X)$. In the first instance a is greater than 1, in which case there is always a global maximum. In the next two instances a is less than 1, so there is a



(a) Global maximum (b) Local maximum (c) No maximum

Figure 3.2: Three instances of the posterior density of the MP2C algorithm. The arrows in (a) and (b) indicate the MAP estimate. In (b) and (c) the 'peaks' at zero are singular.

singularity at zero but only in one case there is a local maximum. The arrows in figures 3.2(a) and 3.2(b) indicate the MAP estimates.

Although it is not evident at first sight (especially in the case of the MMSE), both the MAP and the MMSE estimators give the well-known Wiener solution for $a = 1$

$$\hat{S} = \frac{X\mathrm{E}[S^2]}{\mathrm{E}[S^2] + \mathrm{E}[N^2]} \tag{3.13}$$

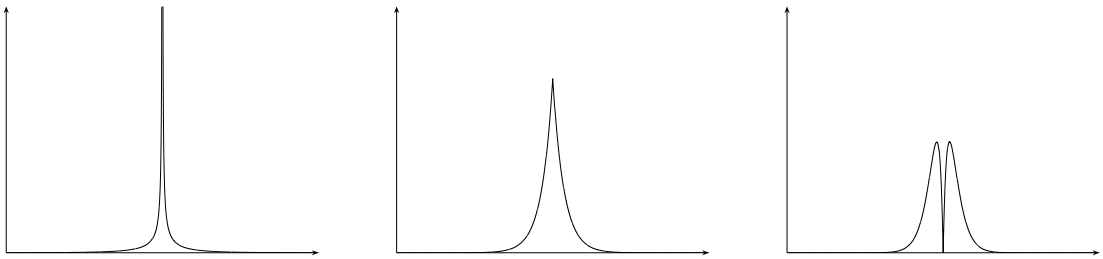
where $\mathrm{E}[S^2]$ is the variance of speech S, which for a general 2 sided Chi pdf, is equal to $\theta a/2$ and in this particular case is $\theta/2$. $\mathrm{E}[N^2]$ is the variance of noise N, which is according to eq. 3.5 is $\mathrm{E}[N^2] = \sigma_N^2$.

3.2.2 2 sided Gamma speech priors

The 2 sided Gamma density function is a generalisation of the Laplacian pdf and is given by

$$p_{2\mathrm{G}}(S) = \frac{1}{2\theta^a \Gamma(a)} \; |S|^{a-1} \; \exp\left[-\frac{|S|}{\theta}\right] \tag{3.14}$$

The 2 sided Gamma pdf is more leptokurtic (has a higher kurtosis, i.e. higher value at zero, longer tails) than the 2 sided Chi pdf for the same value of a. The case $a = 1$ yields the Laplacian pdf. Some plots for characteristic values of a are shown in figure 3.3.



(a) $a = 0.1$ (b) $a = 1$ (c) $a = 2$

Figure 3.3: 2 sided Gamma pdf's for different values of a.

3.2.2.1 MMSE estimator (MS 2G algorithm)

To obtain the MMSE estimator we need to substitute to eq. 3.8 the expression for the likelihood (eq. 3.5) and the Gamma prior, which is given by eq. 3.14. The resulting estimator is given by (see appendix A.4)

$$\hat{S} = a\sigma_N \frac{\exp\left[\frac{\zeta_1^2}{4}\right] D_{-a-1}(\zeta_1) - \exp\left[\frac{\zeta_2^2}{4}\right] D_{-a-1}(\zeta_2)}{\exp\left[\frac{\zeta_1^2}{4}\right] D_{-a}(\zeta_1) + \exp\left[\frac{\zeta_2^2}{4}\right] D_{-a}(\zeta_2)} \tag{3.15}$$

where $\zeta_1 = \frac{\sigma_N}{\theta} - \frac{X}{\sigma_N}, \zeta_2 = \frac{\sigma_N}{\theta} + \frac{X}{\sigma_N}$

In order to express the above estimator as a gain for the noisy coefficients we should first note that the expression for the a priori SNR is now $\xi = \theta^2 a(a+1)/\sigma_N^2$ (see §4.3.2), while the a posteriori SNR is again $\gamma_2 = X^2/\sigma_N^2$. The resulting expression is

$$\hat{S} = X \left[\frac{a\,\mathrm{sgn}(X)}{\sqrt{\gamma_2}} \frac{\exp\left[\frac{\eta_1^2}{4}\right] D_{-a-1}(\eta_1) - \exp\left[\frac{\eta_2^2}{4}\right] D_{-a-1}(\eta_2)}{\exp\left[\frac{\eta_1^2}{4}\right] D_{-a}(\eta_1) + \exp\left[\frac{\eta_2^2}{4}\right] D_{-a}(\eta_2)} \right] \tag{3.16}$$

where $\eta_1 = \frac{\sqrt{a(a+1)}}{\sqrt{\xi}} - \mathrm{sgn}(X)\sqrt{\gamma_2}, \quad \eta_2 = \frac{\sqrt{a(a+1)}}{\sqrt{\xi}} + \mathrm{sgn}(X)\sqrt{\gamma_2}$

The MS2G algorithm with $a = 0.5$ and $a = 1$ has been also proposed by Martin [72].

3.2.2.2 MAP estimator (MP2G algorithm)

The MAP estimator for the 2 sided Gamma priors can be obtained in the same way the corresponding estimator for the 2 sided Chi priors was found. It therefore suffices to find the maximum of $\ln(p(X|S)p(S))$ where $p(X|S)$ is again given by eq. 3.5 and $p(S)$ by eq. 3.14. The resulting estimator is (see appendix A.5)

$$\hat{S} = \zeta + \text{sgn}(X)\left[\zeta^2 + (a-1)\sigma_N^2\right]^{1/2} \qquad \text{where} \quad \zeta = \frac{X}{2} - \text{sgn}(X)\frac{\sigma_N^2}{2\theta} \tag{3.17}$$

The expression of the above estimator as a gain for the noisy coefficients is given below. The expressions for the a priori and the a posteriori SNR's are the same as in §3.2.2.1.

$$\hat{S} = X\left[\eta + \text{sgn}(X)\left[\eta^2 + \frac{a-1}{\gamma_2}\right]^{1/2}\right] \qquad \text{where} \quad \eta = \frac{1}{2} - \frac{1}{2}\sqrt{\frac{a(a+1)}{\xi\gamma_2}} \tag{3.18}$$

When $a < 1$ the 2 sided Gamma density (eq. 3.14) and subsequently the posterior density, given in eq. 3.3, have a singularity at zero. As we discussed in §3.2.1.2 and in order to avoid using the global maximum, which is always at zero, the strategy we follow is to use the local maximum provided by eq. 3.17 when it exists and suppress X by a fixed amount (i.e. 50 dB) when it does not. If we also observe the form of the posterior density (eq. A.29) we can see that the value of S which maximises the posterior density must have the same sign as X. It is possible however, that the expression in eq. 3.17 yields a negative solution for a positive X and vice versa. This is not acceptable and in these cases X is again suppressed by a fixed amount.

No instances of the MP2G algorithm have been found previously in the literature.

3.3 Amplitude algorithms

In the previous section we presented methods for estimating the Re and Im parts of the clean speech STFT coefficients in every frequency bin given the noisy observations. An alternative option is to estimate the amplitude and the phase of the clean speech frequency bins instead, which generates a whole new family of algorithms. In

practice, it is sufficient to estimate the amplitude only and then combine it with the noisy speech phase to create the enhanced speech waveform. That is because it has been widely argued that the perception of speech is phase insensitive [73], [98] and moreover, Ephraim and Malah [31] showed that the optimal estimate for the clean speech phase is the noisy speech phase itself. This property gives the amplitude estimation methods an advantage compared to their DFT coefficients counterparts, which is that the number of data points that need to be estimated is halved.

The STFT coefficients of the noisy speech, the clean speech and the noise in terms of their amplitude and phase are denoted as $\mathbf{X} \equiv R\exp[\mathbf{i}\psi]$, $\mathbf{S} \equiv A\exp[\mathbf{i}\phi]$, and $\mathbf{N} \equiv B\exp[\mathbf{i}\omega]$. The estimation problem can then be formulated as follows: we are trying to find an estimate of the clean speech amplitude A given the noisy speech amplitude R and phase ψ. Recall from §3.2 that in order to derive both the MMSE and the MAP estimators, the calculation of the posterior density $p(A|R,\psi)$ is first necessary. This can be written as

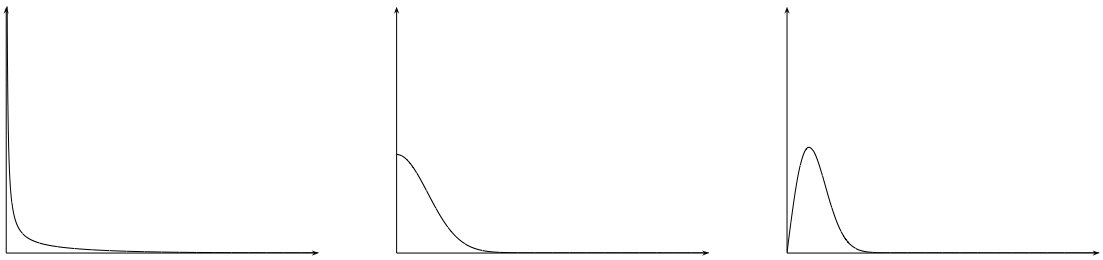
$$\begin{aligned} p(A|R,\psi) &= \frac{p(R,\psi|A)p(A)}{\displaystyle\int_0^\infty p(R,\psi|A)p(A)\ dA} \\ &= \frac{\displaystyle\int_0^{2\pi} p(R,\psi|A,\phi)p(A)p(\phi)\ d\phi}{\displaystyle\int_0^\infty\int_0^{2\pi} p(R,\psi|A,\phi)p(A)p(\phi)\ dA\,d\phi} \end{aligned} \tag{3.19}$$

In the above equation note that $p(A)$ and $p(\phi)$ are factorised, which stems from the assumption that A and ϕ are independent. Simulation results also confirm that the distribution of the clean speech phase is uniform; hence we can replace $p(\phi)$ with $1/2\pi$.

The density function of R and ψ conditioned on A and ϕ is given by (see appendix A.1)

$$p(R,\psi|A,\phi) = \frac{R}{2\pi\sigma_N^2}\ \exp\left[-\frac{R^2+A^2-2RA\cos(\psi-\phi)}{2\sigma_N^2}\right] \tag{3.20}$$

We proceed with the derivation of the MMSE and MAP estimators for different families of speech amplitude priors.



(a) $a = 0.1$ (b) $a = 1$ (c) $a = 2$

Figure 3.4: 1 sided Chi pdf's for different values of a.

3.3.1 1 sided Chi speech priors

The 1 sided Chi density function is the 1 sided version of the pdf described in §3.2.1 and its functional form is given by

$$p_{1\mathrm{C}}(A) = \frac{2}{\theta^{a/2}\Gamma(a/2)}\ A^{a-1}\ \exp\left[-\frac{A^2}{\theta}\right],\ \text{with } A \geq 0 \tag{3.21}$$

For $a = 2$ the above density yields the Rayleigh pdf, while for $a = 1$ one obtains the half Gaussian. Some of its characteristic instances can be seen in figure 3.4. Let us now present the expressions for the MMSE and the MAP estimators.

3.3.1.1 MMSE estimator (MS 1C algorithm)

The MMSE estimator of the clean speech amplitude A given the noisy speech amplitude R and phase ψ is given by

$$\begin{aligned}\hat{A} = \mathrm{E}[A|R,\psi] &= \int_0^\infty A p(A|R,\psi)\ dA \\ &= \frac{\displaystyle\int_0^\infty \int_0^{2\pi} A p(R,\psi|A,\phi) p(A) p(\phi)\ d\phi\, dA}{\displaystyle\int_0^\infty \int_0^{2\pi} p(R,\psi|A,\phi) p(A) p(\phi)\ d\phi\, dA}\end{aligned} \tag{3.22}$$

Substitution of $p(R,\psi|A,\phi)$ and $p(A)$ from eq. 3.20 and 3.21 respectively and the assumption of a uniform phase distribution ($p(\phi) = \frac{1}{2\pi}$) yields (see appendix A.6)

$$\hat{A} = \sqrt{2\sigma_N^2\,\zeta}\ \frac{\Gamma(\frac{a+1}{2})}{\Gamma(\frac{a}{2})}\ \frac{{}_1F_1(\frac{a+1}{2}, 1, \frac{R^2}{2\sigma_N^2}\zeta)}{{}_1F_1(\frac{a}{2}, 1, \frac{R^2}{2\sigma_N^2}\zeta)} \qquad \text{where} \quad \zeta = \frac{\theta}{\theta + 2\sigma_N^2} \tag{3.23}$$

${}_1F_1(\alpha, \beta, z)$ is the Confluent Hypergeometric Function (eq. 9.210.1, [42]). The calculation of ${}_1F_1(\alpha, \beta, z)$ was performed with the mchgm.m routine provided in [8]. To alleviate the numerical problems that occur in the evaluation of the confluent hypergeometric function for large values of its input arguments (typically for $z > 700$) the asymptotic expansions given in eq. 13.5.1 in [1] were used, producing numerically stable results for all input ranges.

The estimator in eq. 3.23 can be expressed as a gain for the noisy coefficients, which is a function of the a priori and the a posteriori SNR's. The relation of the a priori SNR ξ to the scale parameter θ is $\xi = \theta a / 4\sigma_N^2$ (see §4.3.2), while the a posteriori SNR is given by $\gamma = R^2/\mathrm{E}[B^2]$ or $\gamma = R^2/2\sigma_N^2$. The estimator can then be written as

$$\hat{A} = R \left[\sqrt{\frac{\eta}{\gamma}} \, \frac{\Gamma(\frac{a+1}{2})}{\Gamma(\frac{a}{2})} \, \frac{{}_1F_1(\frac{a+1}{2}, 1, \gamma\eta)}{{}_1F_1(\frac{a}{2}, 1, \gamma\eta)} \right] \qquad \text{where} \quad \eta = \frac{\xi}{\xi + a/2} \tag{3.24}$$

The estimator in 3.23 was derived by Loizou [65] from a perceptually motivated point of view. Additionally, the above estimator with $a = 2$ (Rayleigh speech prior) is equivalent to the well known Ephraim-Malah MMSE-STSA algorithm [31].

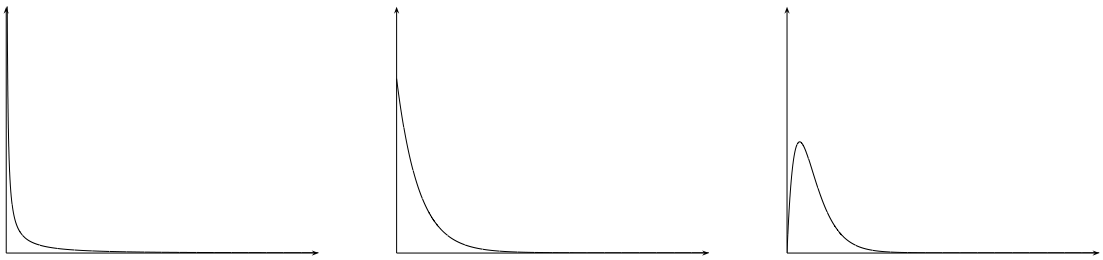
3.3.1.2 MAP estimator (MP 1C algorithm)

The MAP estimator can be found by maximising with respect to A the posterior density $p(A|R, \psi)$. Since the denominator in the expression for the posterior density in eq. 3.19 is not a function of A it suffices to maximise the numerator only, or its logarithm, as this simplifies the calculations significantly; thus

$$\hat{A} = \arg\max_A \; \ln \left(\int_0^{2\pi} p(R, \psi|A, \phi) p(A) p(\phi) \; d\phi \right) \tag{3.25}$$

Substituting $p(R, \psi|A, \phi)$ and $p(A)$ from 3.20 and 3.21 and $p(\phi) = \frac{1}{2\pi}$ yields (see appendix A.7)

$$\hat{A} = \zeta \frac{R}{2} + \left[\left(\zeta \frac{R}{2} \right)^2 + (a - 1.5) \, \sigma_N^2 \, \zeta \right]^{1/2} \qquad \text{where} \quad \zeta = \frac{\theta}{\theta + 2\sigma_N^2} \tag{3.26}$$



(a) $a = 0.1$ (b) $a = 1$ (c) $a = 2$

Figure 3.5: 1 sided Gamma pdf's for different values of a.

The above expression as a gain for the noisy coefficients is

$$\hat{A} = R\left[\frac{\eta}{2} + \left[\left(\frac{\eta}{2}\right)^2 + (a-1.5)\frac{\eta}{2\gamma}\right]^{1/2}\right] \qquad \text{where} \quad \eta = \frac{\xi}{\xi + a/2} \tag{3.27}$$

For $a < 1.5$, the global maximum of the posterior distribution is always at zero, because the posterior density has a singularity at that point. In a similar fashion to the DFT MAP estimators, for $a < 1.5$ we use the local maximum when it exists (when the argument of the square root in eq. 3.26 is positive) and we suppress R by a fixed amount (i.e. 50 dB) when it does not.

Two instances of the MP1C algorithm can be found in the literature: the first is by Wolfe and Godsill [99] using $a = 2$ and the second is by Dat et al. [24] for $a = 1$.

3.3.2 1 sided Gamma speech priors

Another family of speech priors is given by the 1 sided Gamma density function, described by equation

$$p_{1G}(A) = \frac{1}{\theta^a \Gamma(a)}\ A^{a-1}\ \exp\left[-\frac{A}{\theta}\right],\ \text{with } A \geq 0 \tag{3.28}$$

The above pdf is the 1 sided variant of the 2 sided Gamma pdf described in §3.2.2. Some of its characteristic instances for various values of the parameter a are shown in figure 3.5. A well known member of this family of density functions is the exponential, which is obtained for $a = 1$. We now proceed with the derivation of the MMSE and the MAP estimators.

3.3.2.1 MMSE estimator (MS 1G algorithm)

The MMSE estimator is obtained by substituting eqs. 3.28 and 3.20 in 3.22. After some algebraic manipulation of the integrals, which is detailed in appendix A.8, the estimator can be written as

$$\hat{A} = \frac{\mathcal{I}_G(a)}{\mathcal{I}_G(a-1)} \tag{3.29}$$

where

$$\mathcal{I}_G(\nu) \equiv \int_0^\infty A^\nu \exp\left[-\frac{A^2}{2\sigma_N^2} - \frac{A}{\theta}\right] I_0\left(\frac{AR}{\sigma_N^2}\right) dA \tag{3.30}$$

The above integral has no analytic solution for $\nu \in (-1, \infty)$, which is the range of interest for our problem. To solve this problem we resort to numerical integration. It turns out that the integrand in $\mathcal{I}_G$ is sufficiently smooth to allow convergence in a few iterations of the Adaptive Lobatto Quadrature [36]. Additionally, the above estimator could be calculated by means of a look up table in a final implementation of the algorithm, in order to reduce the computational cost imposed by the numerical integration.

No instances of this algorithm have been reported in the literature.

3.3.2.2 MAP estimator (MP 1G algorithm)

The MAP estimator can be found by maximising the expression in (3.25), where the likelihood is again given in (3.20), the phase density is $p(\phi) = \frac{1}{2\pi}$ and the Gamma speech prior is given in (3.28). The resulting estimator is (see appendix A.9)

$$\hat{A} = \zeta + \left[\zeta^2 + (a-1.5)\sigma_N^2\right]^{1/2} \qquad \text{where} \quad \zeta = \frac{R}{2} - \frac{\sigma_N^2}{2\theta} \tag{3.31}$$

For the 1 sided Gamma priors the relation between ξ and θ is $\xi = \theta^2 a(a+1)/2\sigma_N^2$ (see §4.3.2), while the a posteriori SNR is $\gamma = R^2/2\sigma_N^2$. The estimator in eq. 3.31 can be written as

$$\hat{A} = R\left[\eta + \left[\eta^2 + \frac{a-1.5}{\gamma}\right]^{1/2}\right] \qquad \text{where} \quad \eta = \frac{1}{2} - \frac{1}{4}\sqrt{\frac{a(a+1)}{\xi\gamma}} \tag{3.32}$$

In accordance with the other MAP estimators presented so far, when $a < 1.5$ the local maximum is used if it exists, while if it does not exist, R is suppressed by a fixed amount (i.e. 50 dB). The existence of the local maximum is determined by the sign of the argument of the square root in eq. 3.31. Additionally, the above estimator can sometimes yield negative estimates when $a < 1.5$. These estimates are not acceptable, as the parameter we are estimating is amplitude and in these cases R is again suppressed by a fixed amount.

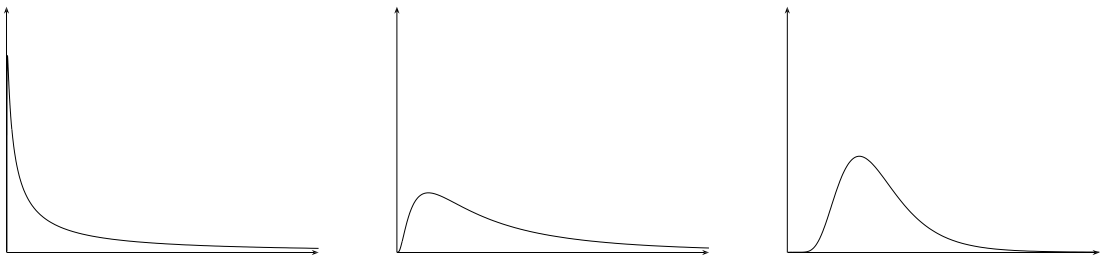
Two instances of this algorithm can be found in the literature: the first is by Lotter and Vary [66], who used $a = 2$ and the second by Dat et al. [24], who used $a = 1.5$.

3.3.3 Lognormal speech priors

Another density function that models very accurately the speech amplitude data is the Lognormal. A random variable has a Lognormal distribution if its logarithmic transformation results in a Gaussian distributed random variable [56]. In other words, if A_{Gauss} is a Gaussian r.v. then $A = \exp(A_{\text{Gauss}})$ follows a Lognormal distribution. Its functional form is given by

$$p_{1\text{L}}(A) = \frac{\sqrt{a}}{\sqrt{\pi}A} \exp\left[-a\left(\ln(A) - \theta\right)^2\right], \text{ with } A \geq 0 \tag{3.33}$$

The similarity with the Gaussian distribution is evident from the above formula, by noting that θ is the mean of the corresponding Gaussian distribution and a is inversely proportional to its variance. The parameter θ can take any value in $\Re$ and controls the scale of the distribution. The parameter a on the other hand, has to be a positive real number and controls the shape of the Lognormal pdf. The effect of the parameter a on the shape of the distribution is illustrated in figure 3.6. A difference between the Lognormal density, compared to the Chi and Gamma, is that its value is zero at the origin (i.e. $p_{1\text{L}}(0) = 0$) for all values of a. Before proceeding to the derivation of the MMSE and MAP estimators with Lognormal speech amplitude priors, we should mention that these two algorithms have never appeared previously in the literature.



(a) $a = 0.1$ (b) $a = 0.5$ (c) $a = 3$

Figure 3.6: Lognormal pdf's for different values of a.

3.3.3.1 MMSE estimator (MS 1L algorithm)

To derive the MMSE estimator with Lognormal speech priors we need to substitute eqs. 3.33 and 3.20 into 3.22. Following a similar procedure as in appendix A.8 the estimator can be reduced to the following form

$$\hat{A} = \frac{\Im_L(0)}{\Im_L(-1)} \tag{3.34}$$

where

$$\Im_L(\nu) \equiv \int_0^\infty A^\nu \exp\left[-\frac{R^2 + A^2}{2\sigma_N^2} - a\,(\ln(A) - \theta)^2\right] I_0\left(\frac{AR}{\sigma_N^2}\right) dA \tag{3.35}$$

The above integral has no analytic solution so numerical integration techniques had to be employed. The calculation was performed with the Adaptive Lobatto Quadrature, in a similar fashion to the amplitude MMSE estimator with Gamma speech priors.

3.3.3.2 MAP estimator (MP 1L algorithm)

To obtain the MAP estimator we need to substitute the expression for the Lognormal prior given in eq. 3.33 into eq. 3.25, together with eq. 3.20 and $p(\phi) = \frac{1}{2\pi}$. The resulting expression is minimised with respect to A. After some simplification, which is detailed in appendix A.10 and after discarding the terms which are constant w.r.t.

A, the expression that has to be maximised is found to be

$$\hat{A} = \arg\max_{A} \left[-\ln(A) - \frac{R^2 + A^2}{2\sigma_N^2} - a\ (\ln(A) - \theta)^2 + \ln\left(I_0\left(\frac{RA}{\sigma_N^2}\right)\right)\right] \quad (3.36)$$

The maximum of this expression cannot be found analytically as equating its first derivative w.r.t. A to zero does not result to an equation whose solution can be obtained in a close form. The maximum is found instead numerically using a quasi-Newton method §10.7 [82].

3.4 Assessment of the independence assumptions

During the development of the estimators in the two previous sections, we assumed that the Re and Im parts and the amplitude and phase of the speech STFT were independent. This assumption simplified significantly the development of the estimators. The dependencies between the elements of the two different representations of the STFT coefficients were reported in [72] to be weak on average, while the amplitude and phase were found to be statistically less dependent than the Re and Im parts. No results however, were given in support of these statements. In this section we quantitatively assess these independence assumptions by measuring the symmetric uncertainty coefficient [82] between the Re and Im and between the amplitude and phase of the clean speech STFT.

The symmetric uncertainty coefficient between two random variables x and y is given by

$$\mathrm{U}(x, y) = 2\frac{\mathrm{H}(x) + \mathrm{H}(y) - \mathrm{H}(x, y)}{\mathrm{H}(x) + \mathrm{H}(y)} \quad (3.37)$$

where H(.) is the entropy of a r.v. The symmetric uncertainty coefficient is a measure of independence between two r.v.'s, which is 0 for independent and 1 for fully dependent r.v.'s. The numerator of eq. 3.37 is the mutual information between x and y, which is denoted by $\mathrm{I}(x, y)$ [59]. For the calculation of the mutual information we used the algorithm proposed in [59]. $\mathrm{H}(x)$ and $\mathrm{H}(y)$ in the denominator are calculated with the same algorithm, exploiting the property $\mathrm{I}(x, x) = \mathrm{H}(x)$ [82].

The clean speech STFT data used in the evaluation was calculated from a clean speech database that consisted of 48 TIMIT sentences uttered by 3 male and 3

female speakers. The total duration of the speech data was 2 minutes and 10 seconds and the sampling frequency 8 KHz. The transformation to the STFT domain was performed with Hamming windows of 256 samples and 75% overlap. For comparison we also calculated the symmetric uncertainty coefficient between the Re and Im and between the amplitude and phase of three test signals. The first of the test signals was a complex Gaussian r.v. with independent Re and Im parts. The second was a complex Laplacian r.v. with independent Re and Im parts, and finally, the third test signal had exponential (1 sided Laplacian) amplitude, which was independent from its uniformly distributed phase. The analytic models of the above signals predict that the amplitude and phase of the first signal are independent, while the amplitude and phase of the second and the Re and Im parts of the third have some dependencies. The symmetric uncertainty coefficient results for the above data are shown in table 3.1.

	Speech	Gaussian Re & Im	Laplacian Re & Im	Exp. Amp. & Unif. Phase
$\mathrm{U}(S_{\mathrm{Re}}, S_{\mathrm{Im}})$	0.03	0	0	0.01
$\mathrm{U}(A, \phi)$	0	0	0.001	0

Table 3.1: Symmetric uncertainty coefficient results for the Re and Im and the amplitude and phase of test and speech STFT data.

Table 3.1 shows that the symmetric uncertainty coefficient results[2] agree with the model predictions for the test data and also indicate that while the amplitude and phase of the speech are independent there are indeed some dependencies between its Re and Im parts. One might have anticipated these results by considering that small shifts in time of the STFT analysis windows would affect the speech phase but not its amplitude. Conversely, a multiplication of the speech time waveform with an arbitrary constant, would have affected its spectral amplitude but not its phase. Both of these examples indicate some form of independence between the speech spectral amplitude and phase.

Despite the fact that some dependencies exist between the Re and Im parts of speech, in the development of the DFT estimators we assume they are independent.

[2]The algorithm did not produce exactly 0 for the zeros shown in table 3.1. It instead produced either negative or very small values ($< 1 \times 10^{-5}$) that varied between realisations for the test data. The authors of [59] state that these cases indicate independent r.v.'s, hence the zeros in the tables. For the used speech data the actual $\mathrm{U}(A, \phi)$ was -6×10^{-6}.

A first reason is the lack of a non Gaussian model that can effectively take these dependencies into account. Additionally, and perhaps more importantly, given the complexity of the estimators of §3.2, any attempt to couple the estimators of the Re and Im parts is likely to cause a further increase in the algorithm's complexity, while any substantial improvement in the performance is dubious as the dependencies between the Re and Im parts are rather weak.

3.5 Summary

In this chapter we derived a number of speech enhancement algorithms that form the backbone of this thesis. We started by formulating the problem of enhancing speech as an estimation problem in the STFT domain. We then derived a framework of STFT speech enhancement algorithms that can be grouped in the following categories: Firstly, according to the STFT feature they estimate, which was either the Re and Im parts (DFT algorithms) or the STFT amplitude (amplitude algorithms). Secondly, according to the estimator they employed, which was the MMSE or the MAP. The final feature of the algorithms were the priors used to model the clean speech samples. For the DFT algorithms, the 2 sided Chi and Gamma priors were used. For the amplitude algorithms, the priors used were the 1 sided Chi and Gamma and the Lognormal.

Two assumptions made during the development of the algorithms were that the Re and Im parts and the amplitude and phase of the speech STFT are independent. These assumptions, which cannot hold simultaneously for other than Gaussian models, were tested in the last section of this chapter. The results showed that although the amplitude and phase are independent, some dependencies exist between the Re and Im parts. Nevertheless, these dependencies were not taken into account in the development of the respective algorithms because they were rather weak, while their incorporation was likely to result in a significant increase in the complexity of the estimators.

Chapter 4

Parameter estimation

The prior densities used in the development of the estimators of chapter 3 have two parameters: the shape parameter a and the scale parameter θ. In the present chapter we shall examine a number of approaches for estimating their values. The estimation methods we discuss can be divided in two categories: the first, is based on fitting the prior densities to a large amount of speech data and extracting the parameter values that provide the optimal fit. The second category includes methods that estimate the parameters adaptively during the enhancement process. Two methods of the first category are discussed in §4.1 and §4.2, while §4.3 and §4.4 discuss two adaptive methods.

The optimal fit of the prior densities to the speech data can be found via the Kullback-Leibler (KL) divergence[1]. Its definition for the discrete case is [60]:

$$\mathrm{KL} = \sum_{m=1}^{N_{\mathrm{bin}}} (p_d(m) - p_s(m)) \ln \left(\frac{p_d(m)}{p_s(m)} \right) \tag{4.1}$$

where $p_d(m)$ is the pdf of the data, calculated from a histogram, and $p_s(m)$ is the speech prior evaluated at the position of the histogram's bins. N_{bin} is the number of bins used for the creation of the histogram. The values of the density function parameters that provide the best fit to the data are those that minimize the KL divergence. The purpose of fitting densities to the data is actually twofold. Apart from extracting values for the parameters, which can subsequently be used with the

[1] A Maximum Likelihood based method was also tried, but resulted in poorer matching of the data distributions.

estimation algorithms, it can also show the appropriateness of the proposed densities for modelling the data.

A first approach in obtaining parameter estimates via the fitting method is to fit the priors to the entire STFT data (full data set) obtained from a large speech database. The results of this method are presented in §4.1. A more refined approach would be the separate fitting of the priors to data extracted from a single frequency bin, thus allowing for variations in the form of the densities that model data from different frequencies. The results of the last approach are shown in §4.2. In both cases however, it must be ensured that the data to which the priors are fitted is scaled, so that it has the same standard deviation with the speech data which is to be enhanced. In the present work, the data used in the evaluation of the speech enhancement algorithms is a subset of that used for fitting the priors, hence the above requirement is met.

Although the above methods can yield estimates for both a and θ, it is beneficial in the implementation of the algorithms to couple one of them with the a priori SNR. The incorporation of the a priori SNR and its estimation with a method such as the DD method [31], is reported to aid the reduction of the background noise level and also to suppress the musical noise artifacts [16]. The a priori SNR is linked by definition to the second moment of the speech samples. Despite the fact that the second moment of all the considered densities is controlled by both a and θ, simulations show that the parameter θ is related to the scale of the density, while the parameter a controls the shape. This can be easily verified by fitting a density function to a random variable multiplied with two different constants, in which case the value of a that provides the best fit remains unaffected, while θ changes according to the multiplying constant. It seems therefore more appropriate that the parameter that is coupled with the a priori SNR is θ. The adaptive estimation of the scale parameter via the a priori SNR and the DD method is discussed in §4.3.

From a Bayesian theoretic point of view, the methods of §4.1 and §4.2 model speech with a long term prior. That is, a prior with fixed values of the scale and shape parameters is employed for modelling the louder and quieter portions of speech as well as the small segments of silence between words. With the introduction of the DD method on the other hand, the priors become local or short term, because their

scale is now a function of the a priori SNR which changes with time.

The estimation of θ via the a priori SNR implies that the use of the estimates of a obtained from long term speech data (§4.1, §4.2) is not justified theoretically. The reason is that the latter methods assume a constant value of θ for the whole duration, which is not the case as θ is adaptively estimated from the a priori SNR. A method for estimating a via the fitting of priors that is compatible with the adaptive estimation model of θ is shown in §4.3.3. Finally, in §4.4 we will present a method for the adaptive estimation of a, which is based on the moment matching method and is also compatible with the estimation of θ from the a priori SNR.

The speech data to which all the priors of this chapter are fitted was taken from the TIMIT database. The data used consisted of 16 male and 16 female speakers, each uttering 8 sentences. After removing the silent frames with a Voice Activity Detector (VAD), the total length of the data was 12.5 minutes. The sampling frequency was 8 KHz, while the STFT transformation was performed with Hamming windows of 256 samples and a 75% overlap. It is conceivable that there might be differences between the distribution of clean speech data, and speech data extracted from real life noisy speech recordings. A possible source of these discrepancies for example might be the Lombard effect. We assume however, that the differences should not be major and proceed with the use of clean speech data, which are significantly easier to obtain.

4.1 Fitting densities to the full data set

We begin by demonstrating the fitting of the proposed densities to the full data set, beginning with the Re and Im parts and then with the amplitude. Figure 4.1(a) shows the histogram of the real part of the full data set and the 2 sided Gamma and Chi densities. The respective histograms for the imaginary parts are essentially identical and are not shown. The parameters used in the densities are those that provided the best fit according to the KL divergence. Figure 4.1(b) shows the central part of figure 4.1(a). Table 4.1 shows the parameter values and the KL divergence values for the Re/Im parts.

As we can see from the above figures and especially from the KL divergence the

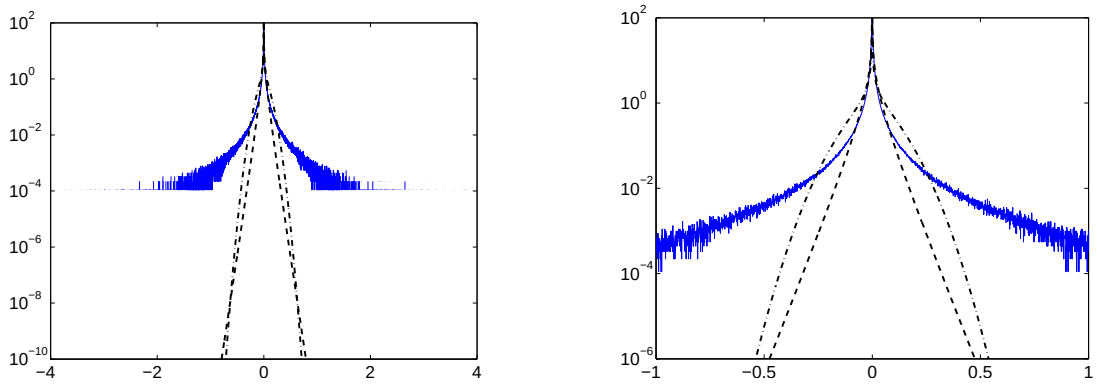


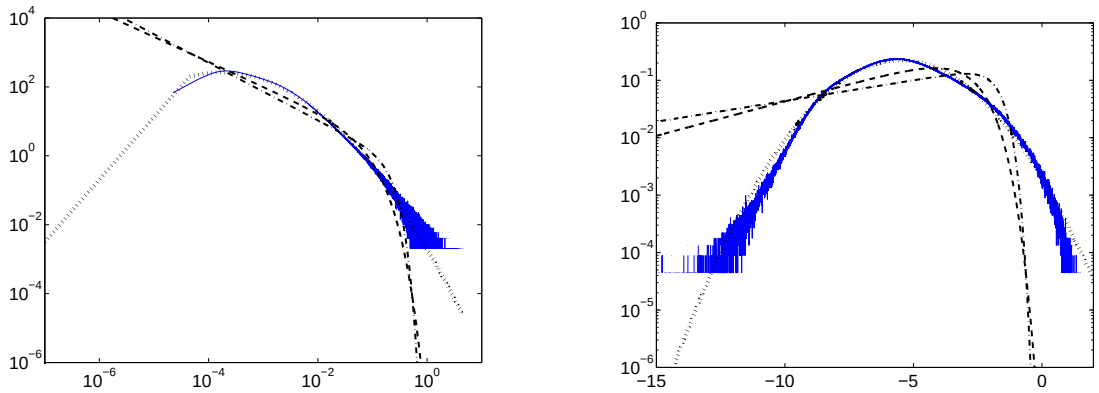
Figure 4.1: (a) Histogram (solid) of the real part of the full data set and fitting of the Gamma (dash) and Chi (dash dot) densities, (b) zoom in the central part of (a).

Density	a	θ	KL
Chi	0.15/0.14	0.024/0.028	669/564
Gamma	0.25/0.24	0.036/0.038	289/228

Table 4.1: Parameter values that minimize the KL divergence when fitting the 2 sided Chi and Gamma densities to the Re/Im parts of the full data set.

Gamma density models the speech data more accurately. In their attempt to capture the large peak at zero however, both distributions underestimate the long tails of the speech data histogram.

Figure 4.2(a) shows the histogram of the spectral amplitude of the full data set and the three densities with parameter values that provide the best fit according to the KL divergence. Because the speech spectral amplitude distribution has a high concentration close to zero, while a few samples have relatively large amplitudes, it is difficult for histograms with a linear data bins segmentation to provide a good resolution for the whole range of values. A remedy for this problem is to calculate the histogram of the logarithm of the speech spectral amplitude instead. This is feasible since amplitude values are always non negative and are practically never zero. Visual evaluation of the fitting of the densities however, requires that they are also transformed into the logarithmic domain. Figure 4.2(b) shows the histogram of the natural logarithm of the speech spectral amplitudes and the corresponding transformed densities. Table 4.2 shows the parameter values that provide the best fit according to the KL divergence. The functional forms of the densities transformed



(a) Amplitude (b) Logarithm of amplitude

Figure 4.2: Histogram (solid) of the amplitude of the full data set and fitting of the Gamma (dash), Chi (dash dot) and Lognormal (dot) densities.

in the logarithmic domain are shown in appendix B.

Density	a	θ	KL
Chi	0.17	0.034	1042
Gamma	0.28	0.056	464
Lognormal	0.16	-5.49	13

Table 4.2: Parameter values that minimize the KL divergence when fitting the 1 sided Chi and Gamma and the Lognormal densities to the amplitude of the full data set.

The results demonstrate clearly that the fitting of the Lognormal density to the data is superior compared to that provided by either the Gamma or the Chi. The Lognormal density has the ability to capture the heavy tails of the speech amplitude data and at the same time model the drop of the distribution as the amplitude values approach zero. The Chi and Gamma densities on the other hand, underestimate the tails of the distribution, and additionally predict that the probability density increases as we move toward zero, which is not in agreement with the evidence provided by the data.

4.2 Fitting densities to each frequency bin

Instead of fitting the densities to data taken from all the frequency bins it is possible to fit the distributions to the data in each frequency bin separately. This approach

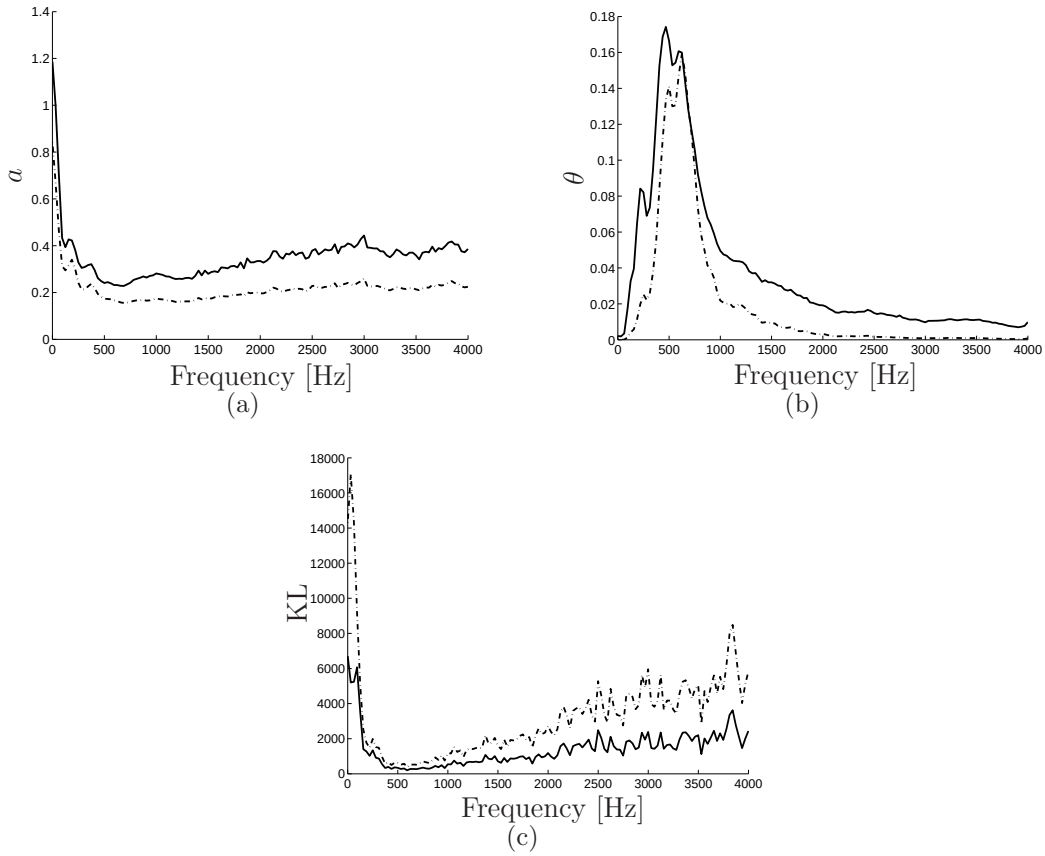


Figure 4.3: Results from fitting the 2 sided Chi (dash-dot) and Gamma (solid) densities to the real part of each frequency bin.

increases the model's flexibility as it allows the shape and scale of the densities to vary with frequency. The results from applying this model to the real part of the STFT coefficients are shown in figure 4.3. Figure 4.3(a) shows the values of a as a function of the frequency, while figures 4.3(b) and 4.3(c) show the values of θ and the KL divergence respectively. Virtually identical results were obtained from the imaginary parts of the data.

The value of a is associated with the kurtosis of the random variable. A r.v with a high value of kurtosis has a large concentration around one point (e.g. the mean or zero), and long tails. Random variables with low kurtosis have a flatter distribution. For all the examined priors small values of a indicate higher kurtosis. The values of a reach a minimum between 0.5-1 KHz, where most of the speech harmonics lie. The data in these frequency bins have a high concentration around zero, when harmonics are absent, while their presence gives large values to the data. Subsequently, the kurtosis increases and the values of a drop. For higher frequencies, where the

amplitude of the harmonics is smaller, the value of a rises slightly.

As mentioned previously, the parameter θ is mainly influenced by the scaling of the random variable, or in other words, its energy. The high values of θ for the frequency range between 0.2 and 1 KHz indicate that most of the speech energy is present there, which is in agreement with the evidence provided by the speech data. Finally, observation of the KL divergence plot shows that the values of the Chi density are 1.5 to 3 times higher than those of the Gamma, which indicates the better fitting of the Gamma prior to the data. This is consistent with the results obtained when data from all the frequencies was used (see table 4.1).

Similar conclusions can be drawn by the examination of the corresponding plots for the amplitude data, which are shown in figure 4.4. The KL divergence plots show that the Lognormal density values are 2-10 times smaller than those of the Gamma and 5-20 times smaller than those of the Chi. This again illustrates that the Lognormal priors can more accurately model the shape of the speech amplitude data distributions.

4.3 Adaptive estimation of the scale parameter.

In this section we will discuss the adaptive estimation of the scale parameter θ via the a priori SNR ξ. We begin by introducing the DD method for the estimation of the a priori SNR and then we show how the latter quantity can be related to the scale parameter θ for the different priors. Finally, we will consider the implications of the adaptive estimation of θ on the estimation of the shape parameter a.

4.3.1 The a priori SNR and its estimation

The a priori SNR ξ was defined by Ephraim and Malah [31] as:

$$\xi(k,l) = \frac{\mathrm{E}[|\mathbf{S}(k,l)|^2]}{\mathrm{E}[|\mathbf{N}(k,l)|^2]} \tag{4.2}$$

where k and l are the frequency and time indices correspondingly. The proportional relation of $\xi(k,l)$ with the second moment $\mathrm{E}[|\mathbf{S}(k,l)|^2]$ shows that it directly controls

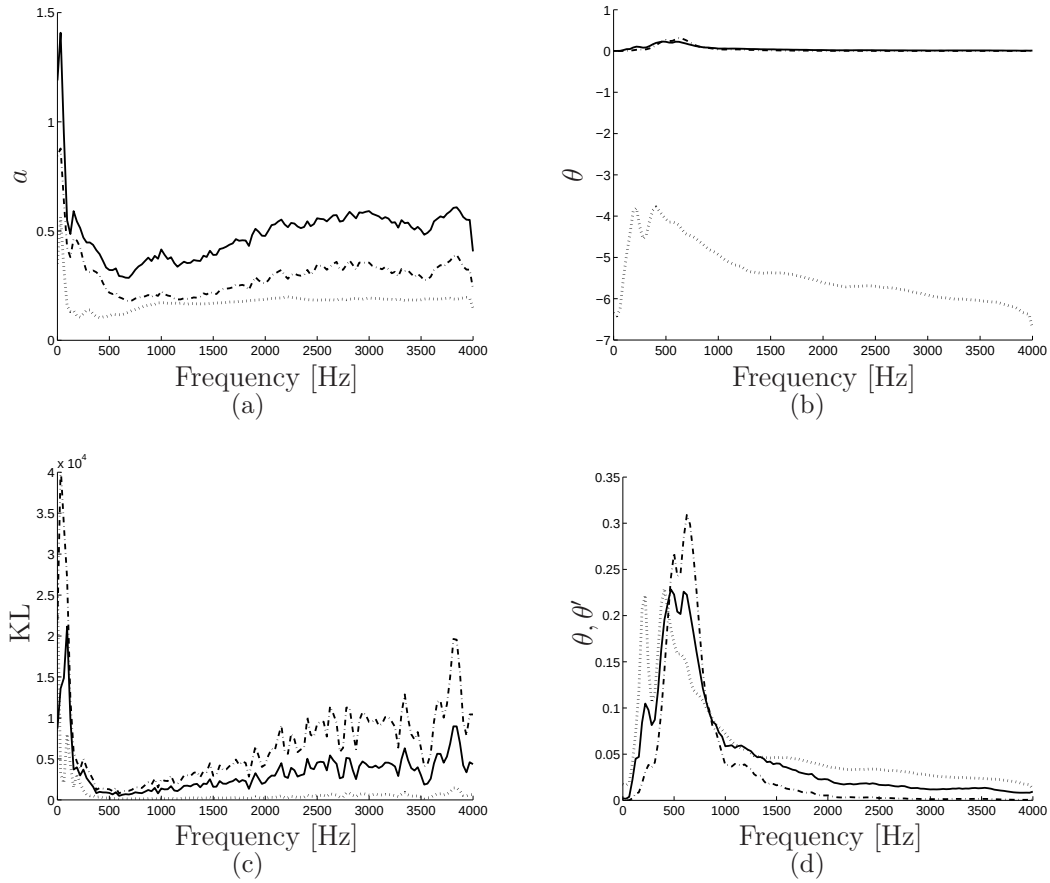


Figure 4.4: Results from fitting the 1 sided Chi (dash-dot), Gamma (solid) and Lognormal (dot) densities to the amplitude of each frequency bin. In (d) the dotted line is a transformation $\theta' = 10\exp(\theta)$ of the values of θ for the Lognormal priors, so that their scaling matches that of the values of θ for the two other priors.

the scaling of the speech prior density, and therefore plays a major part in the estimation process.

The DD method for the estimation of the a priori SNR, which was proposed in [31], is based on the definition of the a priori SNR (eq. 4.2) and on its relation with the a posteriori SNR $\gamma(k,l)$. The definition of the latter is

$$\gamma(k,l) = \frac{|\mathbf{X}(k,l)|^2}{\mathrm{E}[|\mathbf{N}(k,l)|^2]} \tag{4.3}$$

while its relation with the a priori SNR is

$$\xi(k,l) = \mathrm{E}[\gamma(k,l) - 1] \tag{4.4}$$

A linear combination of eqs. 4.2 and 4.4 can be written as

$$\xi(k,l) = \mathrm{E}\left[\alpha\frac{|\mathbf{S}(k,l)|^2}{\mathrm{E}[|\mathbf{N}(k,l)|^2]} + (1-\alpha)(\gamma(k,l)-1)\right] \tag{4.5}$$

The DD estimator, which is based on the above expression, is

$$\hat{\xi}(k,l) = \alpha\frac{|\hat{\mathbf{S}}(k,l-1)|^2}{\mathrm{E}[|\mathbf{N}(k,l-1)|^2]} + (1-\alpha)\max\left[\frac{|\mathbf{X}(k,l)|^2}{\mathrm{E}[|\mathbf{N}(k,l)|^2]} - 1, 0\right] \tag{4.6}$$

The DD estimator was obtained by dropping the expectation operator in eq. 4.5 and using the estimated amplitude from frame $l-1$ instead of the amplitude of frame l. Additionally, the max[.,.] operator ensures the estimator's positiveness.

The advantage of the DD method is that it aids the elimination of the musical noise. The mechanism by which this is achieved is documented by Cappé [16]. Its main attributes are that during speech absence the a priori SNR is a highly smoothed version of the a posteriori SNR, while when speech is present the a priori SNR follows the a posteriori SNR with a delay of 1 frame.

Alternative methods for the estimation of the a priori SNR can also be found in the literature [20–22, 46]. These methods attempt to address the delay in the response of the DD method in an increase of the a priori SNR, and the one frame delay in the periods of speech presence. These methods however, are computationally more complex and they do not share the simplicity in the implementation of the DD method.

4.3.2 Relation of the scale parameter to the a priori SNR

The scale parameter θ can be related to the a priori SNR via its relation with the second moment of $\mathbf{S}$ for each prior. Dropping the time frequency indices for notational simplicity and denoting $\mathrm{E}[|\mathbf{N}|^2]$ as $2\sigma_N^2$, the second moment of the complex STFT coefficient $\mathbf{S}$ can be written as $\mathrm{E}[|\mathbf{S}|^2] = 2\sigma_N^2\xi$. Assuming that the second moments of the Re and Im parts of speech are equal to $\mathrm{E}[S^2]$, that is $\mathrm{E}[S^2] \equiv \mathrm{E}[S_{\mathrm{Re}}^2] = \mathrm{E}[S_{\mathrm{Im}}^2]$, it will then also hold that $\mathrm{E}[S^2] = \mathrm{E}[|\mathbf{S}|^2]/2$. The expressions for the second moments of the 2 sided Gamma and Chi priors are $\mathrm{E}[S^2] = \theta^2 a(a+1)$ and $\mathrm{E}[S^2] = \theta a/2$ respectively. For the 1 sided priors, the expressions for the second moment $\mathrm{E}[A^2]$,

for which holds that $\mathrm{E}[A^2] \equiv \mathrm{E}[|\mathbf{S}|^2]$, are $\mathrm{E}[A^2] = \theta^2 a(a+1)$ for the Gamma, $\mathrm{E}[A^2] = \theta a/2$ for the Chi and $\mathrm{E}[A^2] = \exp(2\theta + a^{-1})$ for the Lognormal. The relations of θ to the a priori SNR ξ for each of the examined priors are summarised in table 4.3.

	Gamma	Chi	Lognormal
2 sided	$\theta^2 = \dfrac{\sigma_N^2 \xi}{a(a+1)}$	$\theta = \dfrac{2\sigma_N^2 \xi}{a}$	-
1 sided	$\theta^2 = \dfrac{2\sigma_N^2 \xi}{a(a+1)}$	$\theta = \dfrac{4\sigma_N^2 \xi}{a}$	$\theta = \dfrac{\ln(2\sigma_N^2 \xi)}{2} - \dfrac{1}{2a}$

Table 4.3: Relation of θ to the a priori SNR ξ for the proposed priors

4.3.3 Fitting densities to narrow variance data

The use of the a priori SNR for the estimation of θ means that the use of long time speech data for fitting the priors and obtaining an estimate of a (§4.1, §4.2) is no longer appropriate. The reason is that the fitting of the speech priors to long time data assumes that the values of a and θ remain constant for the whole duration, which is clearly not the case when the a priori SNR estimates and subsequently the values of θ change with time. To overcome this problem it has been proposed to examine the distribution of speech data from all frequency bins that correspond to a narrow a priori SNR interval. This method has been considered in [66, 67, 72]. In this section we implement the above method and evaluate its results.

The extraction of speech data from narrow a priori SNR intervals is performed using the following procedure: white Gaussian noise is added to the clean speech at a high input segmental SNR[2], e.g. 50 dB. The noise is added to ensure finite values for the a priori SNR. The actual value of the input segmental SNR is not important as long as it is sufficiently high. The input segmental SNR has to be sufficiently high to ensure that the weaker speech components do not get buried in noise and the extraction of an accurate estimate of their a priori SNR is possible. The noisy signal is then enhanced with the Ephraim-Malah algorithm [31] (MS1C algorithm with $a = 2$), which returns an a priori SNR value for each sample of the clean speech STFT. The

[2]For a definition of the segmental SNR see §5.2

DD method smoothing parameter α was set to 0.99. The proposed speech priors were then fitted to data that had a priori SNR values in a narrow interval (1 dB). This interval had to be in a relatively high SNR range, otherwise the data that belonged to it corresponded to noise rather than speech. In our simulation we found that the weaker speech components had an a priori SNR of approximately 20 dB, given an input segmental SNR of 50 dB.

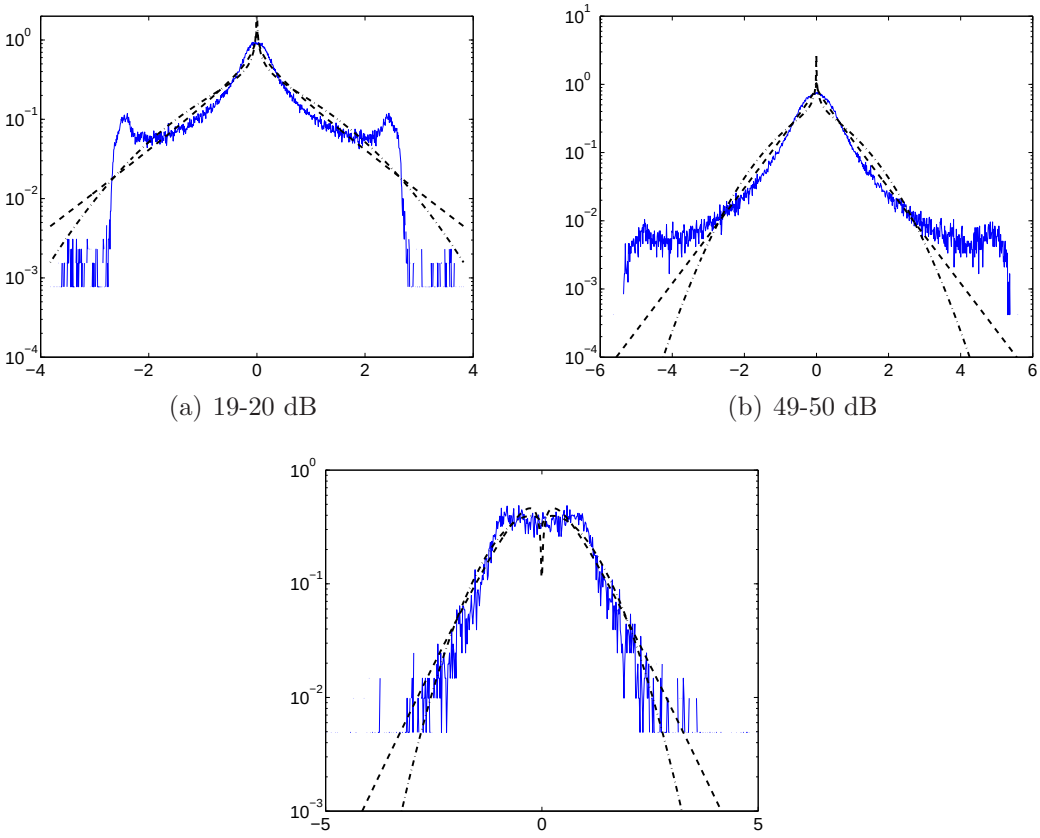
In the following we present results from fitting the proposed priors to data from three intervals, i.e. 19-20, 49-50 and 79-80 dB. The first interval consisted of weak speech components like consonants, while the last interval corresponded to high amplitude data, typically found in the harmonics of the pitch period of vowels. Figure 4.5 shows the histograms of the real parts of the DFT data from the three intervals and the fitted densities. Table 4.4 shows the KL divergence values that corresponded to the best fit that could be achieved with each density for the Re and Im parts and the respective values of the priors' parameters.

Density	Interval dB	a	θ	KL
Chi	19-20	0.58/0.58	2.96/2.89	20/22
-\-	49-50	0.88/0.81	2.02/2.16	5/5
-\-	79-80	1.30/1.32	1.59/1.54	13/13
Gamma	19-20	0.87/0.88	0.69/0.67	7/7
-\-	49-50	1.15/1.19	0.61/0.59	1/1
-\-	79-80	1.68/1.74	0.54/0.52	22/22

Table 4.4: Parameter values that minimize the KL divergence when fitting the 2 sided Chi and Gamma densities to the Re/Im part of data from a narrow variance interval.

The KL divergence values for each case reveal once again that the Gamma density provides a more accurate fit than the Chi. It is worth also noting that as the SNR interval moves to higher ranges, the value of the parameter a increases; that is, the value of the distributions at zero decreases and the tails decay faster, i.e. the distribution becomes more platykurtic.

Figure 4.6 shows the fitting of the amplitude priors to data from narrow variance intervals and table 4.5 shows the corresponding values. When the SNR interval is in a low range the Lognormal density fits the data better, capturing more accurately the data distribution for both small and large values. As the a priori SNR interval

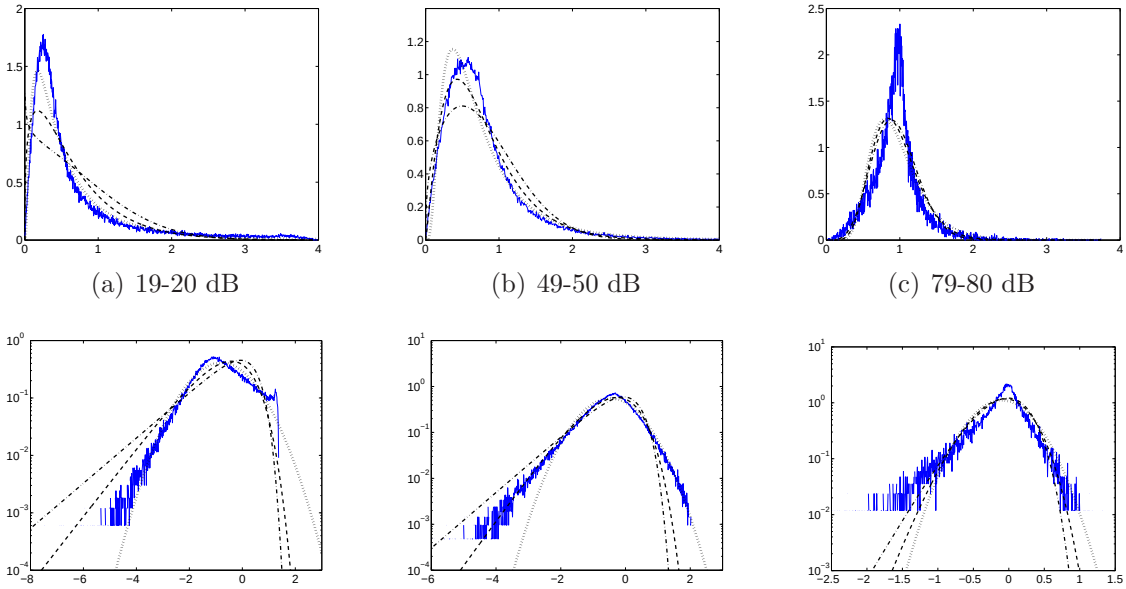


(a) 19-20 dB

(b) 49-50 dB

(c) 79-80 dB

Figure 4.5: Fitting the 2 sided Gamma (dash) and Chi (dash dot) densities to DFT data from a narrow variance interval.



(a) 19-20 dB (b) 49-50 dB (c) 79-80 dB

Figure 4.6: Fitting the 1 sided Gamma (dash), Chi (dash dot) and Lognormal (dot) densities to amplitude data from a narrow variance interval. Upper row presents the histograms of the amplitude and the lower row the histograms of the natural logarithm of the amplitude.

moves to higher ranges, the Lognormal density underestimates the probability of small values, which is better captured by the Gamma density for 49-50 dB and by the Chi for 79-80 dB. The tails of the distribution however, are better modelled by the Lognormal density in all three cases.

The main drawback of the parameter estimation approach described in this section is that the distribution of the data and subsequently the estimated values of a showed a strong dependence on the a priori SNR interval. This does not allow us to extract a single value of a that optimally fits the data. Furthermore, a mapping between the value of the a priori SNR interval and the values of a cannot be obtained, as such a mapping depends on the global SNR, which is not known in general.

In view of the above observations, the strategy we adopt, when the scale parameter θ is estimated adaptively, is to use a range of different values of a with the algorithms and evaluate their performance as a function of the shape parameter a. In this way, the optimal values of a are decided a posteriori, based on the results of the speech enhancement algorithms. Ephraim and Malah [31] also proposed an a posteriori evaluation of their statistical model, in order to sidestep the problems arising from the inaccessibility of the true statistical model of speech.

Density	Interval dB	a	θ	KL
Chi	19-20	0.91	1.96	106
-\-	49-50	1.39	1.28	25
-\-	79-80	4.98	0.39	51
Gamma	19-20	1.36	0.49	49
-\-	49-50	2.25	0.34	7
-\-	79-80	8.88	0.11	57
Lognormal	19-20	0.53	-0.80	14
-\-	49-50	0.96	-0.47	8
-\-	79-80	3.85	-0.10	78

Table 4.5: Parameter values that minimize the KL divergence when fitting the 1 sided Chi, Gamma and Lognormal densities to amplitude data from a narrow variance interval.

4.4 Adaptive estimation of the shape parameter

In this section we present an adaptive method for the estimation of the shape parameter a, which is based on moment matching. The proposed method is also found in [24], although a different strategy for the estimation of moments was employed in that work. An application of a similar cumulant based method to an image estimation problem method was proposed in [34]. A maximum likelihood method for finding estimates of a is described in [89], but requires a significantly greater amount of computation and the availability of clean speech samples. The proposed moment matching method on the other hand, is simple in its implementation and can be applied directly on the noisy samples. In the following, we derive the expressions for the estimators of the shape parameter a for the different priors, as a function of the second and fourth moments of the noise, speech and noisy speech signals. The estimation of the various moments and the evaluation of the proposed adaptive method for the estimation of a, will be detailed in §5.5.

4.4.1 Estimation of a for the 2 sided priors

4.4.1.1 2 sided Chi priors

Given the model $X = S + N$ for the Re (or Im) part of the noisy speech, clean speech and noise coefficients, the fourth moment of the noisy speech can be written

as:

$$\mathrm{E}[X^4] = \mathrm{E}[S^4] + 6\mathrm{E}[S^2]\mathrm{E}[N^2] + \mathrm{E}[N^4] \tag{4.7}$$

The Gaussian noise model of eq. 3.5 yields the following expressions for the second and fourth moments

$$\mathrm{E}[N^2] = \sigma_N^2, \qquad \text{and} \qquad \mathrm{E}[N^4] = 3\sigma_N^4$$

The fourth moment of N can then be expressed in terms of $\mathrm{E}[N^2]$ as

$$\mathrm{E}[N^4] = 3\left(\mathrm{E}[N^2]\right)^2 \tag{4.8}$$

Similarly, the corresponding moments of the 2 sided Chi pdf are:

$$\mathrm{E}[S^2] = \theta a/2, \qquad \text{and} \qquad \mathrm{E}[S^4] = \theta^2 a(a+2)/4$$

The fourth moment in terms of the second can be expressed as

$$\mathrm{E}[S^4] = \frac{a+2}{a}\left(\mathrm{E}[S^2]\right)^2 \tag{4.9}$$

Substituting eqs. 4.8 and 4.9 in 4.7 we obtain:

$$\mathrm{E}[X^4] = \frac{a+2}{a}\left(\mathrm{E}[S^2]\right)^2 + 6\mathrm{E}[S^2]\mathrm{E}[N^2] + 3\left(\mathrm{E}[N^2]\right)^2$$

or:

$$\frac{a+2}{a} = \frac{\mathrm{E}[X^4] - 6\mathrm{E}[S^2]\mathrm{E}[N^2] - 3\left(\mathrm{E}[N^2]\right)^2}{\left(\mathrm{E}[S^2]\right)^2} = \kappa_2 \tag{4.10}$$

Therefore, the estimator of the shape parameter is

$$\hat{a} = \frac{2}{\kappa_2 - 1} \tag{4.11}$$

In eq. 4.10 κ_2 can be recognised as the *kurtosis* of the Re (Im) parts of clean speech, which is defined as $\kappa_2 \equiv \mathrm{E}[S^4]/\left(\mathrm{E}[S^2]\right)^2$. Note that as the kurtosis tends to infinity, a tends to zero and the priors are getting narrower with longer tails. As the kurtosis approaches 1, which is its theoretical lower limit, the value of a tends to infinity.

4.4.1.2 2 sided Gamma priors

An estimate for the a parameter can be obtained in a similar way as in the previous section. The corresponding moments for the Gamma prior are:

$$\mathrm{E}[S^2] = \theta^2 a(a+1), \qquad \text{and} \qquad \mathrm{E}[S^4] = \theta^4 a(a+1)(a+2)(a+3)$$

Therefore, the relation of the fourth moment to the second is

$$\mathrm{E}[S^4] = \frac{(a+2)(a+3)}{a(a+1)} \left(\mathrm{E}[S^2]\right)^2 \tag{4.12}$$

Following the same procedure as in §4.4.1.1 we have:

$$\frac{(a+2)(a+3)}{a(a+1)} = \frac{\mathrm{E}[X^4] - 6\mathrm{E}[S^2]\mathrm{E}[N^2] - 3\left(\mathrm{E}[N^2]\right)^2}{\left(\mathrm{E}[S^2]\right)^2} = \kappa_2 \tag{4.13}$$

Solving the quadratic equation we have:

$$a = \frac{5 - \kappa_2 \pm \sqrt{(5-\kappa_2)^2 - 24(1-\kappa_2)}}{2(\kappa_2 - 1)} \tag{4.14}$$

Finally, simplifying the argument of the square root, the estimator of a becomes

$$\hat{a} = \frac{5 - \kappa_2 + \sqrt{\kappa_2^2 + 14\kappa_2 + 1}}{2\kappa_2 - 2} \tag{4.15}$$

In the solution of the quadratic equation, the root with the $(+)$ is selected because for $\kappa_2 > 1$, which are the acceptable values for the kurtosis, the root with the $(-)$ is negative as is evident from eq. 4.14.

4.4.2 Estimation of a for the 1 sided priors

4.4.2.1 1 sided Chi priors

Given the model for the complex STFT coefficients $\mathbf{X} = \mathbf{S} + \mathbf{N}$ the fourth moment of the noisy speech spectral amplitude can be written as:

$$\mathrm{E}[R^4] = \mathrm{E}[A^4] + 4\mathrm{E}[A^2]\mathrm{E}[B^2] + \mathrm{E}[B^4] \tag{4.16}$$

where R, A and B are the amplitudes of the noisy speech, the clean speech and the noise respectively. Based on the Gaussian noise model of eq. 3.5, the second and fourth moments of the noise spectral amplitude are given by

$$\mathrm{E}[B^2] = 2\sigma_N^2, \qquad \text{and} \qquad \mathrm{E}[B^4] = 8\sigma_N^4$$

and the second and fourth moments are related by

$$\mathrm{E}[B^4] = 2\left(\mathrm{E}[B^2]\right)^2 \tag{4.17}$$

The corresponding moments for the speech spectral amplitude for the Chi prior density are:

$$\mathrm{E}[A^2] = \theta a/2, \qquad \text{and} \qquad \mathrm{E}[A^4] = \theta^2 a(a+2)/4$$

and subsequently

$$\mathrm{E}[A^4] = \frac{a+2}{a}\left(\mathrm{E}[A^2]\right)^2 \tag{4.18}$$

Substituting eqs. 4.17 and 4.18 in 4.16 we have

$$\frac{(a+2)}{a} = \frac{\mathrm{E}[R^4] - 4\mathrm{E}[A^2]\mathrm{E}[B^2] - 2\left(\mathrm{E}[B^2]\right)^2}{\left(\mathrm{E}[A^2]\right)^2} = \kappa_1 \tag{4.19}$$

The estimator of a then reads

$$\hat{a} = \frac{2}{\kappa_1 - 1} \tag{4.20}$$

where κ_1 is the kurtosis of the clean speech amplitude, defined as $\kappa_1 \equiv \mathrm{E}[A^4]/\mathrm{E}[A^2]^2$.

Note that the form of eq. 4.20 is the same as eq. 4.11, which is a consequence of the second and fourth moments being the same for the 1 sided and the 2 sided Chi pdf's.

4.4.2.2 1 sided Gamma priors

The procedure for obtaining the estimates of a is identical to that of §4.4.2.1, except for the expressions of the speech prior moments. For the 1 sided Gamma prior these are:

$$\mathrm{E}[A^2] = \theta^2 a(a+1), \qquad \text{and} \qquad \mathrm{E}[A^4] = \theta^4 a(a+1)(a+2)(a+3)$$

and

$$\mathrm{E}[A^4] = \frac{(a+2)(a+3)}{a(a+1)} \left(\mathrm{E}[A^2]\right)^2 \tag{4.21}$$

Following the same steps as in §4.4.2.1 we have:

$$\frac{(a+2)(a+3)}{a(a+1)} = \frac{\mathrm{E}[R^4] - 4\mathrm{E}[A^2]\mathrm{E}[B^2] - 2\left(\mathrm{E}[B^2]\right)^2}{\left(\mathrm{E}[A^2]\right)^2} = \kappa_1 \tag{4.22}$$

Or finally, solving the quadratic equation w.r.t a:

$$\hat{a} = \frac{5 - \kappa_1 + \sqrt{\kappa_1^2 + 14\kappa_1 + 1}}{2\kappa_1 - 2} \tag{4.23}$$

The valid root from the solution of the quadratic equation is the one with the (+) for the same reasons as those stated in §4.4.1.2. Note again that eq. 4.23 is identical to eq. 4.15, which is the consequence of the second and fourth raw moments of the 1 sided and 2 sided Gamma density functions being identical.

4.4.2.3 Lognormal priors

The expressions for the second and the fourth moments of the Lognormal priors are [56]:

$$\mathrm{E}[A^2] = \exp\left(2\theta + a^{-1}\right), \qquad \text{and} \qquad \mathrm{E}[A^4] = \exp\left(4\theta + 4a^{-1}\right)$$

and the two moments are related by

$$\mathrm{E}[A^4] = \exp\left(2a^{-1}\right) \left(\mathrm{E}[A^2]\right)^2 \tag{4.24}$$

Following the same procedure as in §4.4.2.1 we can show that

$$\exp(2a^{-1}) = \frac{\mathrm{E}[R^4] - 4\mathrm{E}[A^2]\mathrm{E}[B^2] - 2\left(\mathrm{E}[B^2]\right)^2}{\left(\mathrm{E}[A^2]\right)^2} = \kappa_1 \tag{4.25}$$

Solving the above equation with respect to a, we have the following expression for the estimator

$$\hat{a} = \frac{2}{\ln(\kappa_1)} \tag{4.26}$$

4.5 Summary

The priors we employ for modelling the speech STFT data have two parameters: the scale parameter θ and the shape parameter a. In this chapter we proposed a number of methods for estimating their values. The proposed methods were grouped in two categories: the first category contains methods that estimate the parameters by fitting the priors to long term speech data, while the second consists of adaptive methods.

The methods that use long term speech data were two: the first method used data from all the available frequency bins, while the second method involved fitting the priors to data from each frequency bin separately. In both cases, the best fit was provided by the Lognormal priors. The Gamma priors offered a somewhat poorer fit and the Chi priors were generally the least successful models. The priors estimated with the above methods can be called long term priors, because long term speech data are used for the estimation of their parameters.

Enhancing speech using fixed values of θ, as estimated from the long term priors, results in musical noise artifacts, as we will show in the next chapter. For this reason we investigated an adaptive method for the estimation of the shape parameter θ, which is based on the DD method for the estimation of the a priori SNR. The DD method is renown for aiding the reduction of the musical noise artifacts, while the priors it defines are short term, as the values of their parameters change during the enhancement of speech.

The selection of an adaptive method for the estimation of the scale parameter implies that the use of long term estimates for the shape parameter a is not justified theoretically. We implemented a method for the estimation of a that is found in the literature and is compatible with the estimation of θ via the DD method. This method estimates a via fitting the priors to data from narrow a priori SNR intervals. We showed that the results of this method are not consistent and depend strongly on the selection of the a priori SNR interval. In view of the shortcomings of this method, in the following chapter we evaluate the performance of the algorithms as a function of the shape parameter a and seek an optimal value based on the results.

Finally, an adaptive method for the estimation of the shape parameter a was also

developed, which was based on moment matching. Expressions for the estimators of a were analytically derived for each of the employed priors, while the results of this method are also evaluated in the next chapter.

Chapter 5

Evaluation

In this chapter we present the results from the evaluation of the of Bayesian algorithms described in chapter 3. The evaluation is based on simulations performed with a number of clean speech phrases, artificially corrupted with additive white Gaussian and car noise, which are then enhanced with the proposed algorithms. The performance of the algorithms is measured using a number of objective measures, while formal and informal listening tests are employed to subjectively assess the quality of the enhanced speech.

Of particular interest in this evaluation, is the effect of the priors' shape parameter a on the quality of the enhanced speech. In §5.3 the performance of the algorithms is evaluated as a function of the shape parameter a, where it is revealed that its value essentially controls the trade off between the musical character of the residual noise and its overall level, while the preservation of the weaker speech spectral components is influenced to some extent. In the same section there is also a discussion on the performance of the algorithms with values extracted with the methods presented in §4.1 - §4.3. In §5.4, optimal values for a that maximise the speech quality are sought, by means of a formal subjective listening test. Finally, the adaptive scheme for the estimation of a presented in §4.4 is evaluated in §5.5. Prior to the presentation of the results however, some details about the specifics of the performed simulations and the employed evaluation measures will be first given in §5.1 and §5.2 .

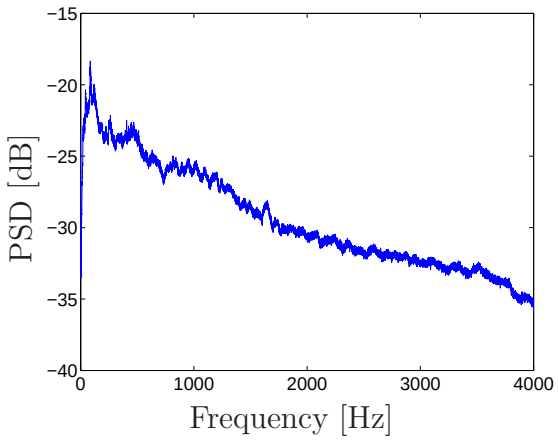


Figure 5.1: Car noise power spectral density.

5.1 Simulation setup

The clean speech database used for the simulations in this chapter is a subset of the database that was used in chapter 4. It comprises of three male and three female speakers, each uttering 8 sentences. The total duration of the database is 2 minutes and 10 seconds and the sampling frequency is 8 KHz. The transformation to the frequency domain was performed using Hamming windows of 256 samples length, overlapped by 75%. The windows were also normalised so that their amplitude when overlapped and added was 1.

The speech phrases were corrupted with white Gaussian and car noise at 0, 10 and 20 dB input Segmental SNR. For these input Segmental SNR levels the corresponding noisy speech PESQ scores were 2.11, 2.80 and 3.46 for the white noise and 2.89, 3.49 and 4.07 for the car noise respectively[1]. The white noise was computer generated, while the car noise was recorded in a car traveling on a motorway at 60 mph. The car noise contained not apparent transients or long term trends, and its power spectrum is shown in figure 5.1. To eliminate the effect of a noise estimation algorithm on the speech enhancement schemes, the noise power was estimated directly from the noise samples, which were known as the mixing of the noise with speech was performed artificially. In practice however, the noise power can be estimated with a noise estimation algorithm, such as those described in chapter 6.

[1]For a definition of Segmental SNR and PESQ see §5.2.

5.2 Methods used for the evaluation of the algorithms

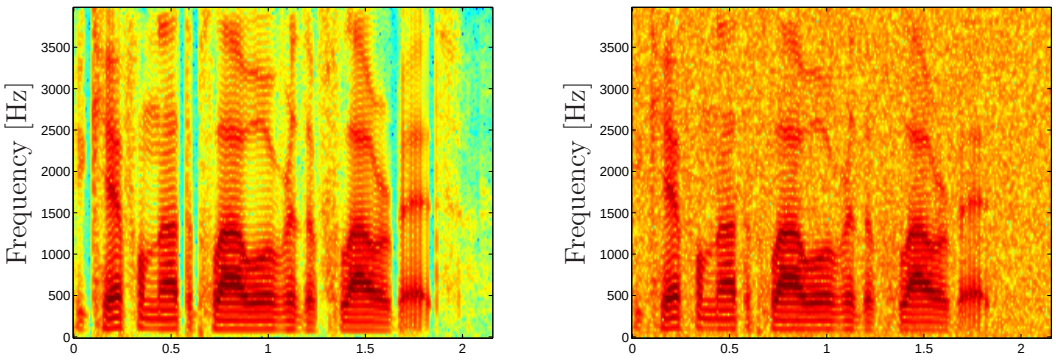
The methods that assess the quality of an enhanced speech utterance can generally be divided into two categories: the subjective and the objective methods. The subjective methods typically involve a panel of listeners who are presented with a set of enhanced speech utterances and subjectively judge their quality, usually based on a predetermined scale. Objective methods on the other hand, are based on a mathematical model, which may or may not try to predict the results of a subjective method. In this work we have used two objective measures: the Segmental Signal to Noise Ratio (SegSNR) and the Perceptual Evaluation of Speech Quality (PESQ).

The SegSNR is an extension of the traditional (or total) SNR and is designed to measure more accurately the quality of the enhanced speech. The Segmental SNR is calculated by finding the logarithm of the SNR in each time frame and then averaging across the frames. Analytically it is given by [25]:

$$\text{SegSNR} = \frac{1}{L}\sum_{l=0}^{L-1} 10\log_{10}\left[\sum_{m=1}^{K}\frac{s^2(Jl+m)}{[s(Jl+m)-\hat{s}(Jl+m)]^2}\right] \tag{5.1}$$

where L is the number of speech frames, K is the number of samples per frame and J is the distance (in samples) between the start points of two consecutive frames. s is the clean and $\hat{s}$ the enhanced speech signal. The motivation for this measure is to emphasize the effect of noise in the low energy speech segments, which are more sensitive to noise compared to the high energy ones. Indeed, a segment with a very low SNR will contribute much more toward the final result in eq. 5.1, because of the addition of the logarithms, whereas with the total SNR the square errors would be summed across the entire waveform. A problem that arises often when the SegSNR is used, is that the existence of silent frames in the signal can produce large negative SNR's, which are not representative of the enhanced speech quality. This problem however, is sidestepped if the silent frames are identified in the clean speech and excluded from the calculation of the SegSNR. This strategy has also been followed in this work, where only the frames that were classified as containing speech, with the aid of a VAD, were used in the calculation of the SegSNR.

The PESQ algorithm [53, 87] is an objective speech quality measure, which has been



(a) Clean Speech (b) Input SegSNR = 10 dB

Figure 5.2: Spectrograms of clean and noisy speech at 10 dB input SegSNR.

approved as the International Telecommunication Union recommendation ITU-T P.862. It is designed to predict the results of a subjective Mean Opinion Score (MOS) test. The scores of the PESQ algorithm lie on a scale from 1 (= bad) to 4.5 (= no distortion). The correlation of the results of the PESQ algorithm with results from subjective MOS tests has been studied in several works [51, 102], where the correlation coefficient was found to be 0.65 in [51] and 0.86 in [102]. Although there is probably still room for improvement in the prediction of MOS results, among the several well known objective speech quality measures evaluated in the above studies, PESQ was the one that presented the highest correlation.

Throughout the presentation of the results, along with the above two objective measures we will give an informal subjective evaluation of the degraded audio samples. The evaluation will be mainly focused on aspects such as the nature of the residual noise (musical vs. broadband) and the amount of speech distortion, which are not always illustrated in the numerical results of the objective measures. A visual supplement will be provided by spectrograms of an enhanced speech segment. The spectrograms of this chapter will correspond to the phrase (*Be careful not to plough over the flower beds*), unless stated otherwise. In order to facilitate a comparison between different algorithms, the spectrograms will also be normalised, so that same colors indicate same spectral amplitude values. The spectrograms of the above phrase prior to noise corruption and mixed with white Gaussian noise at 10 dB input SegSNR are shown in figure 5.2 for reference purposes.

A popular visualisation of an algorithm's properties is given by its suppression

curves. These are plots of the suppression the algorithm applies (in dB) as a function of some of its input parameters. Some of the properties that are illustrated in the suppression curves include the transparency (0 dB suppression) of an algorithm in high input SNR conditions and some indications about the quality of the residual noise (musical or broadband). A number of suppression curves will be given for each algorithm and for some key values of the prior density function parameter a, so that their shape for the whole range of values of a should be easily inferred. The suppression curves will be shown as a function of the a priori and the a posteriori SNR, as it is customary (e.g. [99]).

5.3 Evaluation of the algorithms as a function of the shape parameter a

In chapter 4 we used a number of methods for estimating values for the priors' parameters a and θ. Some of these methods were based on fitting the priors to long term speech data (§4.1, §4.2), while others were estimating the values of the parameters adaptively (§4.3, §4.4). When the values obtained with the methods of the first category are used, the resulting speech suffers from high levels of musical residual noise. On the other hand, the adaptive estimation of the scale parameter via the DD method manages to eliminate the musical noise to a large extent. For this reason, all the algorithms in the remainder of this chapter will use the DD method for the estimation of the scale parameter. The smoothing factor α of the DD method is set to 0.99. Additionally, a lower limit is also set for the a priori SNR at -25 dB, as this was reported to further aid the reduction of musical noise [16]. A comparison of the results obtained with fixed (long term) and adaptive values of the scale parameter is given in appendix C.

In §4.3.3 we also examined a method for the estimation of the shape parameter a, which was based on fitting the priors to data that belonged to narrow a priori SNR intervals and was compatible with the DD method. Unfortunately, this method failed to result in a consistent set of values for a. For this reason, in the current section we examine the performance of the algorithms for a range of values of a, which includes those values that produce the highest scores in the objective measures. In

the following section we attempt to extract optimal values for the shape parameter by means of a formal listening test. The adaptive method for the estimation of a, which was presented in §4.4 and is also compatible with the estimation of θ via the DD method, will be evaluated in §5.5. We now proceed with the evaluation of the performance of the proposed algorithms as a function of the priors' shape parameter a.

Because the signals that are enhanced using the same estimator are acoustically similar, and in order to facilitate the presentation of the results, we will separately discuss the performance of the algorithms according to the estimator used and provide a comparison at the end. We begin with the algorithms that use the MAP estimator.

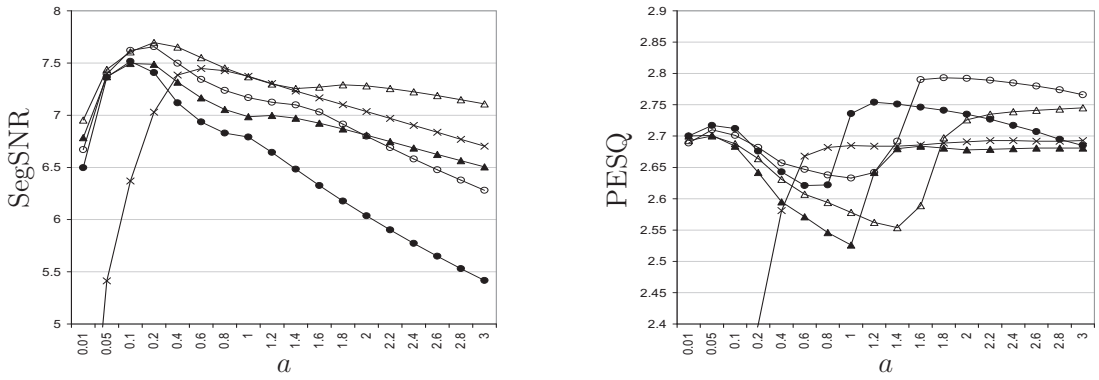
5.3.1 MAP estimator algorithms

Figure 5.3 shows the SegSNR and PESQ scores for the MAP algorithms that estimate either the Re and Im parts or the amplitude of the STFT using the Chi, Gamma and Lognormal speech priors. The results correspond to 3 different input SegSNR values. The corrupting noise is Gaussian and white. Figure 5.4 shows the respective results for car noise. The suppression curves of the above algorithms are shown in figures 5.5 and 5.6.

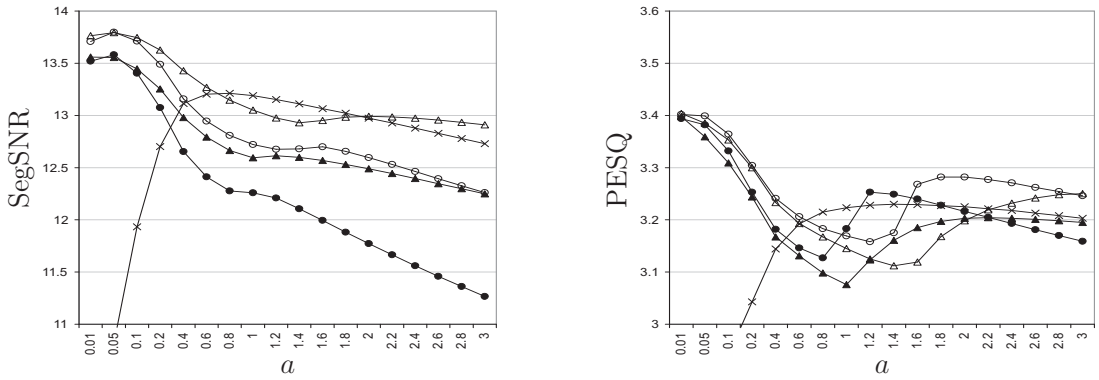
The behaviour of the MAP algorithms that use the Chi or Gamma priors is somewhat different compared to the MAP algorithm that uses the Lognormal priors. The differences arise when the MAP algorithms with the Chi and Gamma priors employ values of a for which the posterior density has a singularity at zero[2] (i.e. $a < 1$ for the DFT MAP algorithms and $a < 1.5$ for their amplitude counterparts). For the above reason we will start the discussion with the MAP algorithms that use the Chi and Gamma priors and the evaluation of the MP1L algorithm will follow.

Evaluation of the MAP algorithms with Chi and Gamma priors In analysing the performance of the MAP algorithms with the Chi and Gamma priors we can

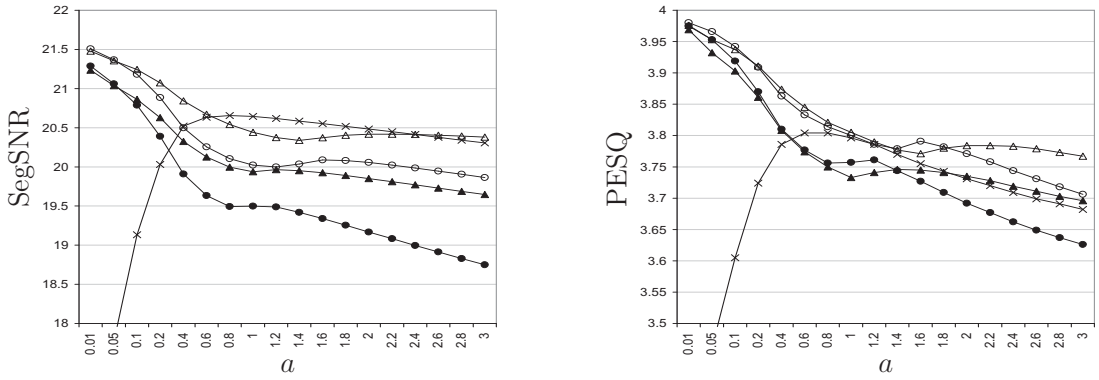
[2]Recall from §3.3.3 that the posterior density $p(A|R,\psi)$ of the MAP algorithm with the Lognormal priors has no singularity at zero for any value of the parameter a.



(a) Input SegSNR = 0 dB, Input PESQ = 2.11

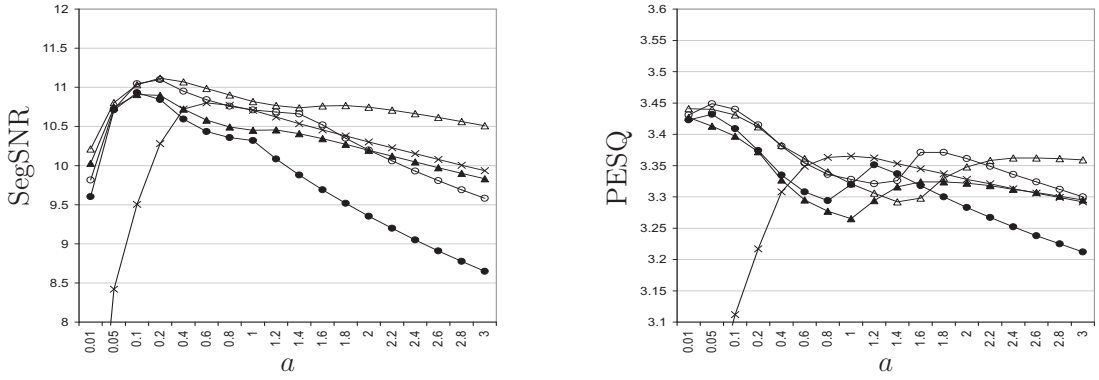


(b) Input SegSNR = 10 dB, Input PESQ = 2.80

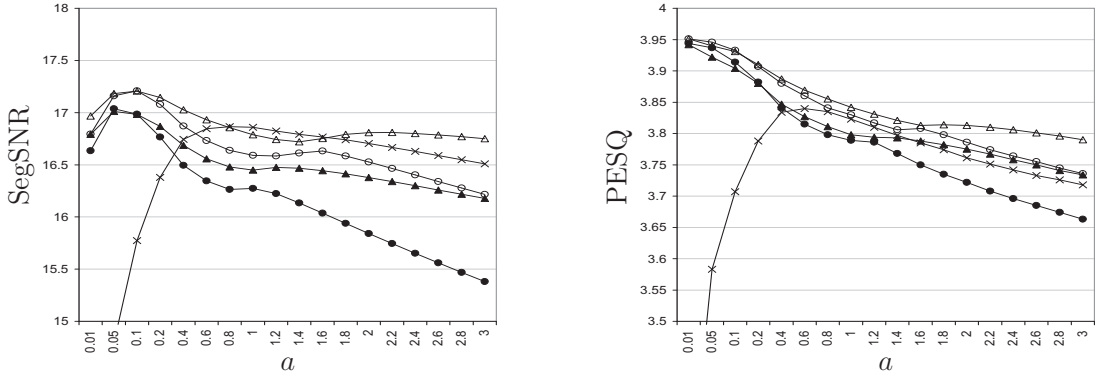


(c) Input SegSNR = 20 dB, Input PESQ = 3.46

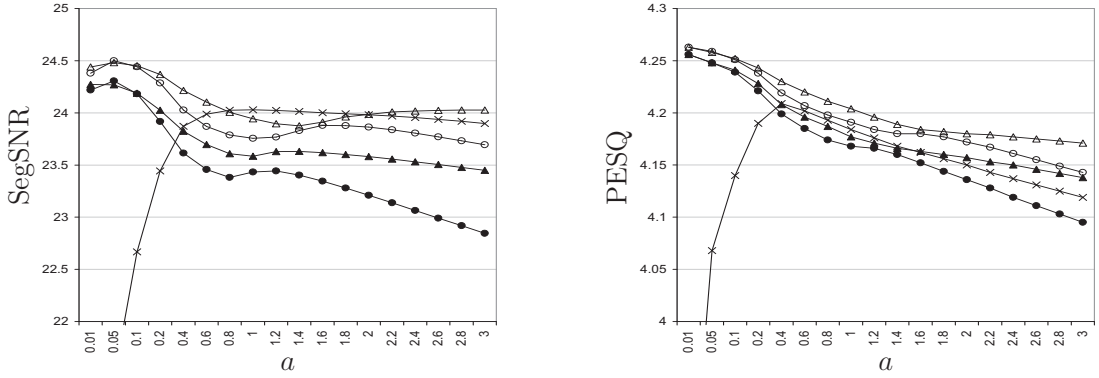
Figure 5.3: SegSNR and PESQ scores for the MAP algorithms for different values of a and input SegSNR's. Speech was corrupted with white Gaussian noise.



(a) Input SegSNR = 0 dB, Input PESQ = 2.89



(b) Input SegSNR = 10 dB, Input PESQ = 3.49



(c) Input SegSNR = 20 dB, Input PESQ = 4.07

Figure 5.4: SegSNR and PESQ scores for the MAP algorithms for different values of a and input SegSNR's. Speech was corrupted with car noise.

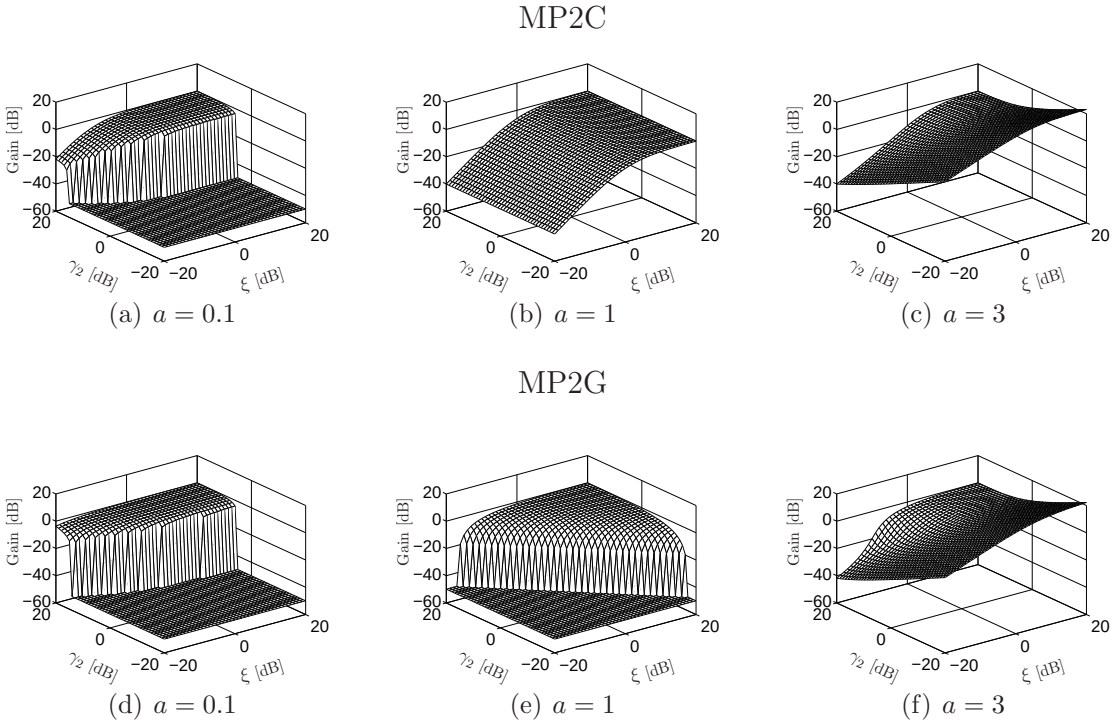


Figure 5.5: Suppression curves of the DFT MAP algorithms for different values of a as a function of the a priori SNR ξ and a posteriori SNR γ_2.

identify two discrete ranges of a, which depend on the existence (first range) or non existence (second range) of the singularity in the posterior distribution. The MAP algorithms with values from the first range preserve adequately the speech components, especially for $a \sim 0.1$. However, the residual noise has a strong musical character. Although the MAP algorithms with values of a from the second range are less successful in recovering the weaker speech components, the residual noise has a more uniform character. The broadband residual noise is also indicated by the 'counter-intuitive' behaviour of the MAP suppression curves (see figures 5.5(c,f) and 5.6(c,f,i)), which show that the suppression increases with increasing values of a posteriori SNR for high values of a[3]. Furthermore, the level of the residual noise, which increases with the value of a, can be adjusted so that the majority of the spurious spectral peaks are masked. On the basis of the uniform character of the residual noise, we prefer values of a from the second range.

Figure 5.7 shows two characteristic instances of the MP1G algorithm with $a = 0.1$ and $a = 3$. In the first case there is a better preservation of the speech spectral

[3]For an explanation of this mechanism see [16].

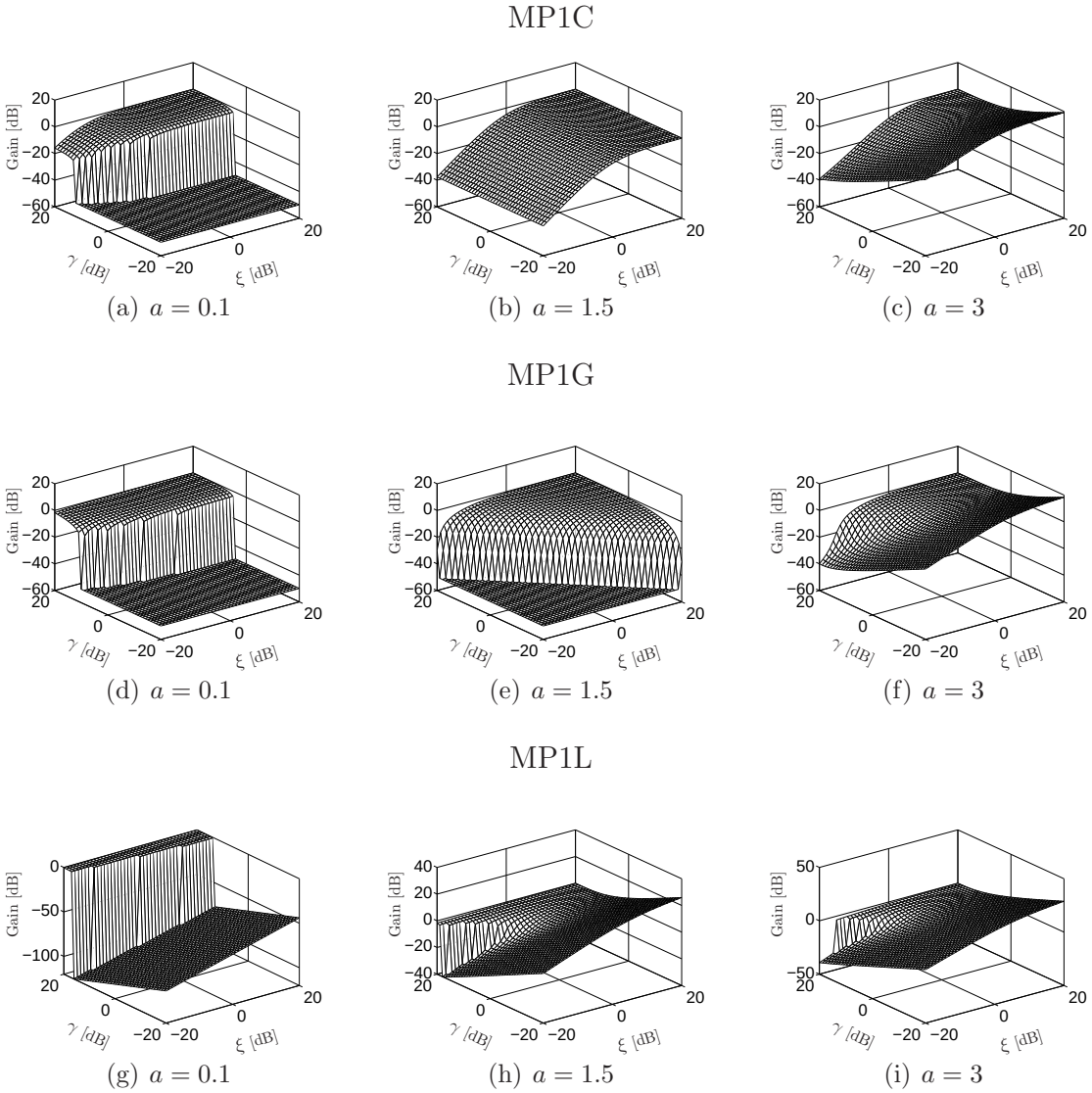
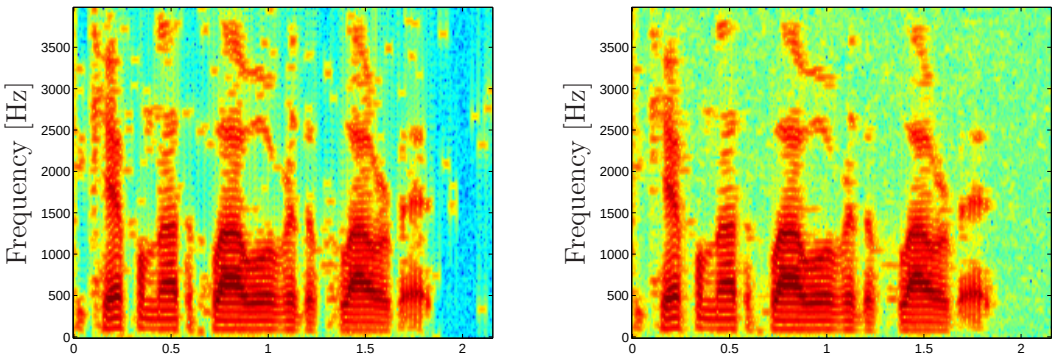


Figure 5.6: Suppression curves of the amplitude MAP algorithms for different values of a as a function of the a priori SNR ξ and a posteriori SNR γ.

components, for instance at 0.5 and 1.5 sec for frequencies above 2.5 KHz. The residual noise however, despite its low level, exhibits a large number of spurious spectral peaks, which are perceived as musical noise.

The MP1L algorithm The behaviour of the MP1L algorithm for values of a larger than 0.4 is similar to that of the remaining MAP algorithms with values of a from the second range. That is, the restoration of the weaker speech components is moderate but the residual noise is uniform. For values of a smaller than 0.4 the MP1L algorithm results in very low residual noise levels but as the value of a drops an increasing number of speech spectral components are also suppressed.



(a) $a = 0.1$ (b) $a = 3$

Figure 5.7: Speech enhanced with the MP1G algorithm for 2 different values of a. Small values of a result in a better preservation of the weaker speech components, while larger values result in uniform residual noise.

This behaviour is reflected in the rapid drop of the objective scores for the MP1L algorithm for a smaller than 0.4 (figures 5.3(e,f), 5.4(e,f)).

To provide a comparison between the examined MAP algorithms we use values of a that result in equal levels of residual noise. In order to obtain these values we concatenate a speech utterance with a segment of silence and enhance the resulting signal with different algorithms, adjusting a so that the output SegSNR's at the silence segment are equal. We find that using the above values of a the resulting signals are very similar acoustically, and their differences can be identified only through careful listening. We proceed with a comparison with respect to the estimated STFT feature and then with a comparison with respect to the employed prior.

Comparison w.r.t the estimated STFT feature A comparison between the DFT and amplitude MAP estimators that use the same priors and values of a that result in equal levels of residual noise reveals that the DFT algorithms slightly underestimate some speech harmonics. Subsequently, the speech enhanced with the amplitude estimators is perceived somewhat louder. The above observation is illustrated in figure 5.8, which shows the SNR for each frame from speech enhanced with the MP1G (continuous line) and the MP2G (dotted line) algorithms, for values of a that resulted in equal levels of residual noise. Note the slightly higher SNR values of the MP1G algorithm at the peaks of speech activity.

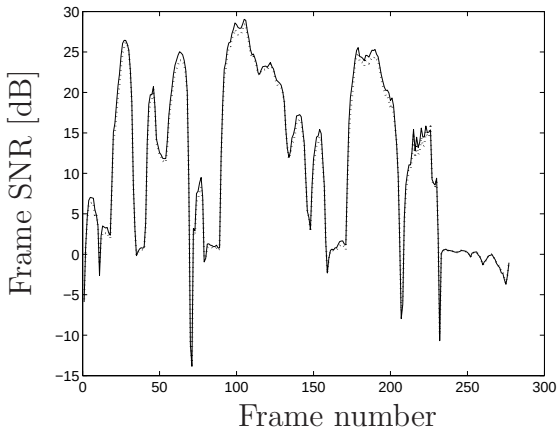


Figure 5.8: Frame SNR as a function of frame number for the MP1G (continuous line) and MP2G (dotted line).

An additional advantage of all the algorithms that work in the amplitude domain is that the total amount of data that needs to be processed is half compared to that of the DFT domain algorithms. The reason is that the amplitude algorithms only estimate the clean speech amplitude, which is then combined with the noisy phase, while the DFT algorithms have to separately estimate the real and the imaginary parts of the STFT.

Comparison w.r.t. the prior Before comparing the three speech priors, we should mention that they all produce speech of very similar quality, when their shape parameter is tuned so as to result in equal levels of residual noise. This ability of the different priors to produce speech of similar quality should be attributed to the flexibility that is provided by their shape parameter a. The differences, which are rather minor, are described in the following. If, according to figures 4.5, 4.6, we classify the three priors according to the length of their tails from the shortest to the longest as Chi, Gamma and Lognormal, we can make the following observations: the use of a prior with shorter tails results in the preservation of a few more weak speech spectral components, at the expense of a larger number of spurious spectral peaks. A longer tailed prior on the other hand, suppresses some of the weaker spectral speech components, but at the same time, the fewer spurious spectral peaks reduce the amount of the perceived speech distortion. Additionally, longer tailed priors have a slightly faster response at the onset of speech after a segment of silence.

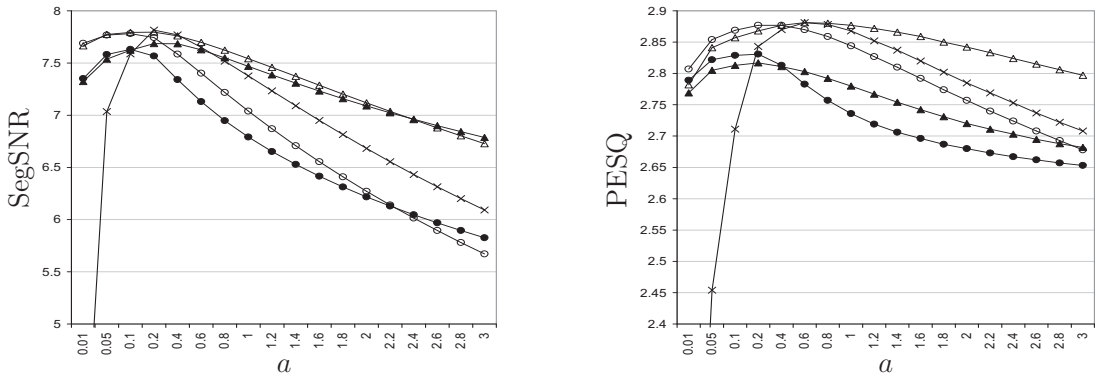
A comment on the approximation of MP1C and MP1G algorithms Among all the algorithms that are examined in this chapter, the amplitude MAP algorithms that use the Chi and Gamma priors are the only approximate estimators, because the Bessel function that appears in the derivation of the respective likelihoods is approximated with eq. A.43 (see appendices A.7, A.9). The same is not true for the MP1L algorithm which is an exact estimator. Enhanced speech of similar quality can be obtained with the three amplitude MAP estimators, while, as we will see in the following section, the same is also true for the three amplitude MMSE estimators (MS1C, MS1G, MS1L), which are all exact (not approximate). The above observation could be an indication that the performance of the exact MP1C and MP1G estimators would not be greatly different from that of the approximate ones, while the latter have the advantage that can be derived in a closed form, which makes them more efficient computationally.

5.3.2 MMSE estimator algorithms

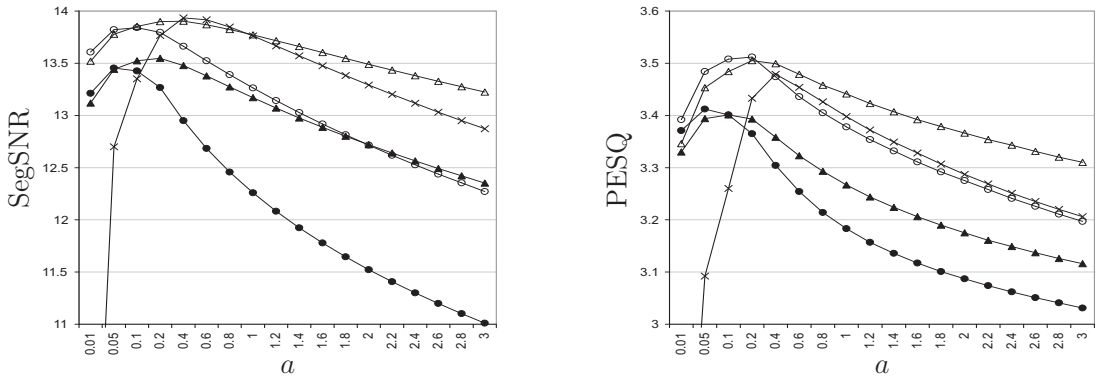
We now present and discuss the results of the algorithms that use the MMSE estimator. Figures 5.9 and 5.10 show the SegSNR and PESQ results for the MMSE algorithms, for different input SegSNR levels and a values. The corrupting noise for the results presented in figure 5.9 was white Gaussian, while the respective results obtained with the car noise are shown in figure 5.10. The suppression curves of the DFT MMSE algorithms for some characteristic values of a are shown in figure 5.11 while figure 5.12 shows the respective suppression curves for the amplitude MMSE algorithms.

General evaluation The MMSE algorithms provide an adequate preservation of the speech spectral components for small values of a (~ 0.2), although the residual noise in this range of the shape parameter has a strong musical character. For increasing values of a the behaviour of the DFT MMSE algorithms is different from that of their amplitude counterparts, so we will discuss them separately.

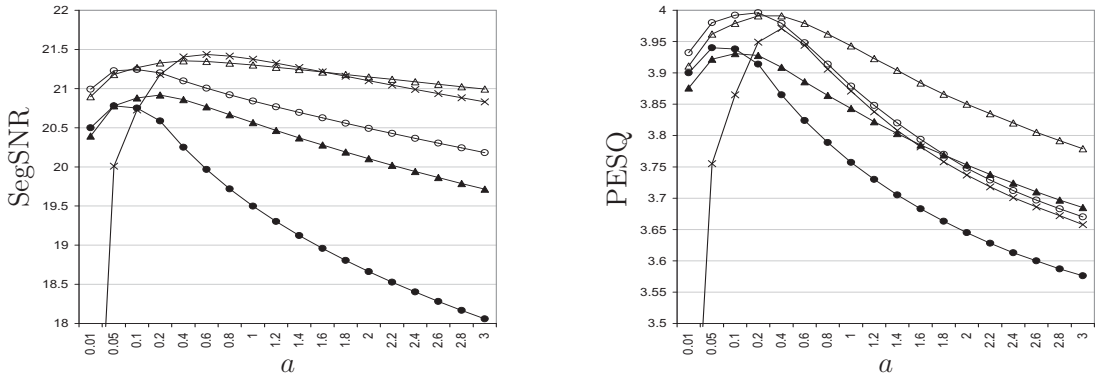
Unlike the MAP algorithms, the residual noise of the DFT MMSE algorithms does not increase with increasing values of a, a fact which is reflected in the almost constant shape of the respective suppression curves (figure 5.11) for small values of



(a) Input SegSNR = 0 dB, Input PESQ = 2.11

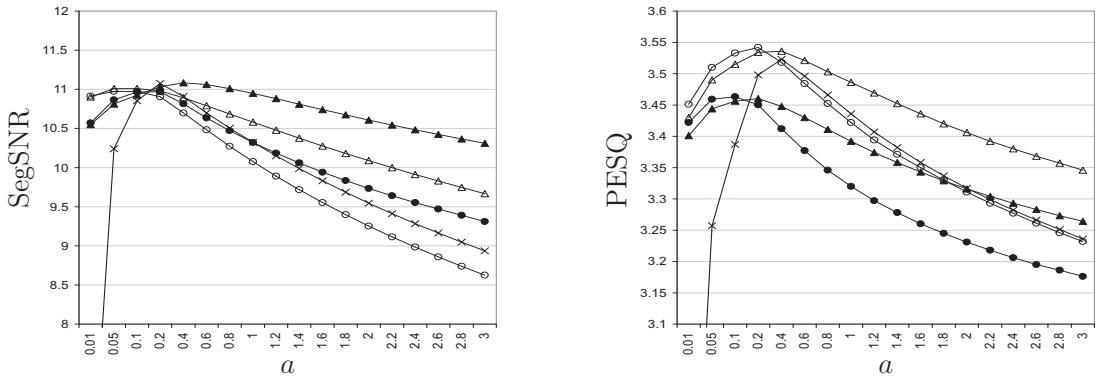


(b) Input SegSNR = 10 dB, Input PESQ = 2.80

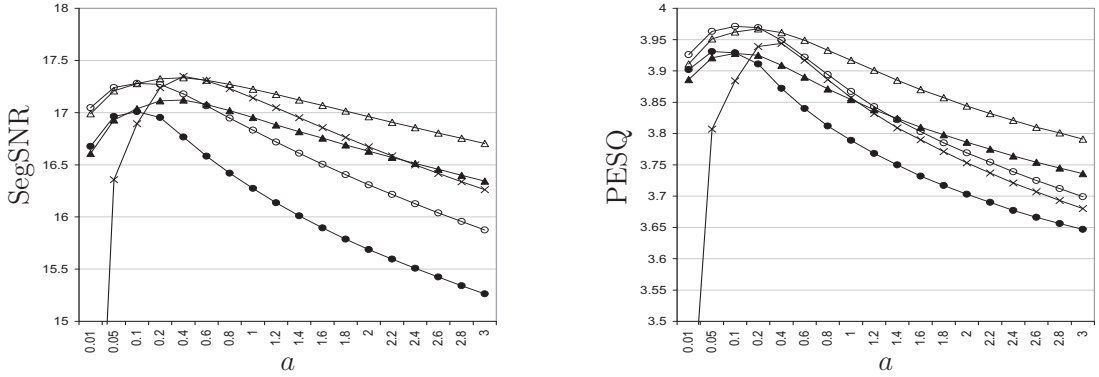


(c) Input SegSNR = 20 dB, Input PESQ = 3.46

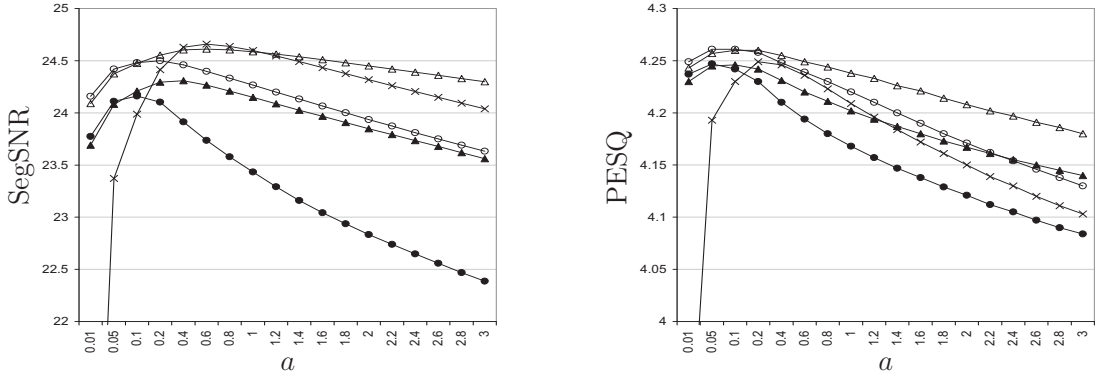
Figure 5.9: SegSNR and PESQ scores for the MMSE algorithms for different values of a and input SegSNR's. Speech was corrupted with white Gaussian noise.



(a) Input SegSNR = 0 dB, Input PESQ = 2.89



(b) Input SegSNR = 10 dB, Input PESQ = 3.49



(c) Input SegSNR = 20 dB, Input PESQ = 4.07

Figure 5.10: SegSNR and PESQ scores for the MMSE algorithms for different values of a and input SegSNR's. Speech was corrupted with car noise.

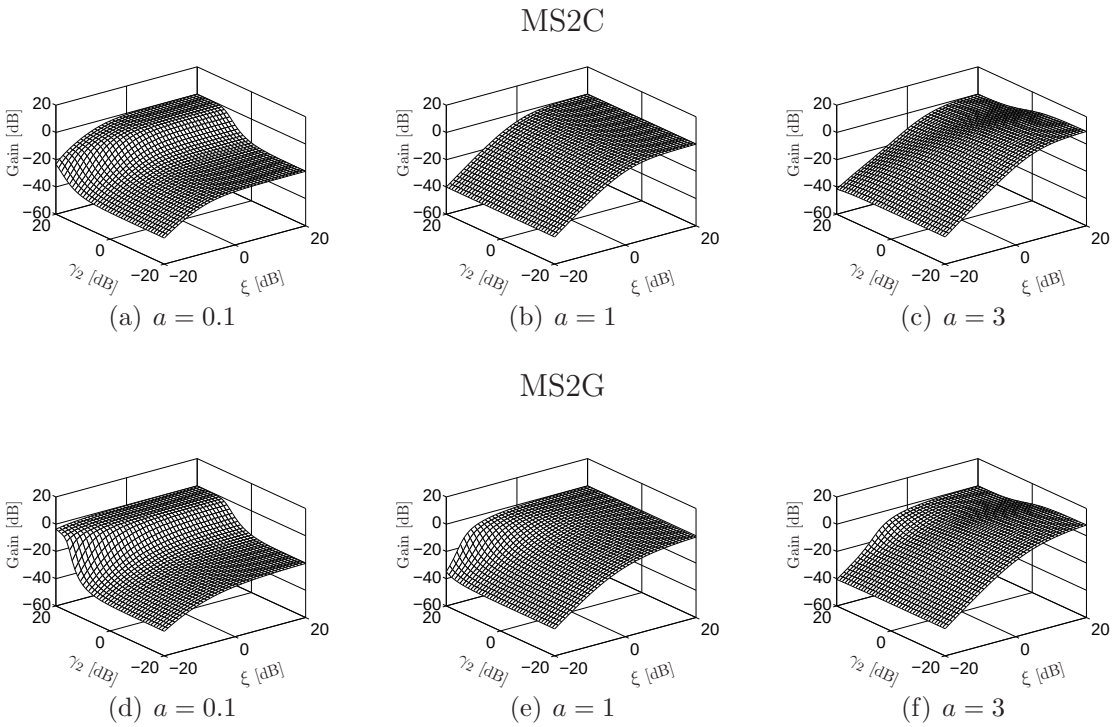


Figure 5.11: Suppression curves of the DFT MMSE algorithms for different values of a as a function of the a priori SNR ξ and a posteriori SNR γ_2.

the a priori SNR and medium to small a posteriori SNR values. Conversely, the level of the spurious spectral peaks decreases, which results in a reduction in the intensity of the musical noise. At the same time however, spectral components that belong to speech are also attenuated, which causes the drop in the objective measures for increasing values of a (figures 5.9, 5.10). The above observations are illustrated in figure 5.13. A discerning characteristic of the DFT MMSE algorithms is that their residual noise has a musical character for all values of the shape parameter a, which can be a fundamental limitation for audio speech enhancement applications.

The amplitude MMSE algorithms on the other hand, do not attenuate the speech spectral components for increasing values of a. They do result however, in an increase in the level of the residual noise, which eventually becomes uniform. The similar shape of the suppression curves of the amplitude MMSE algorithms (figure 5.12) with those of the MAP for low a priori SNR conditions and large values of a gives an indication about the uniform character of the residual noise[4]. Two characteristic instances of the MS1C algorithm for small and large values of a that

[4]See also the third paragraph of §5.3.1

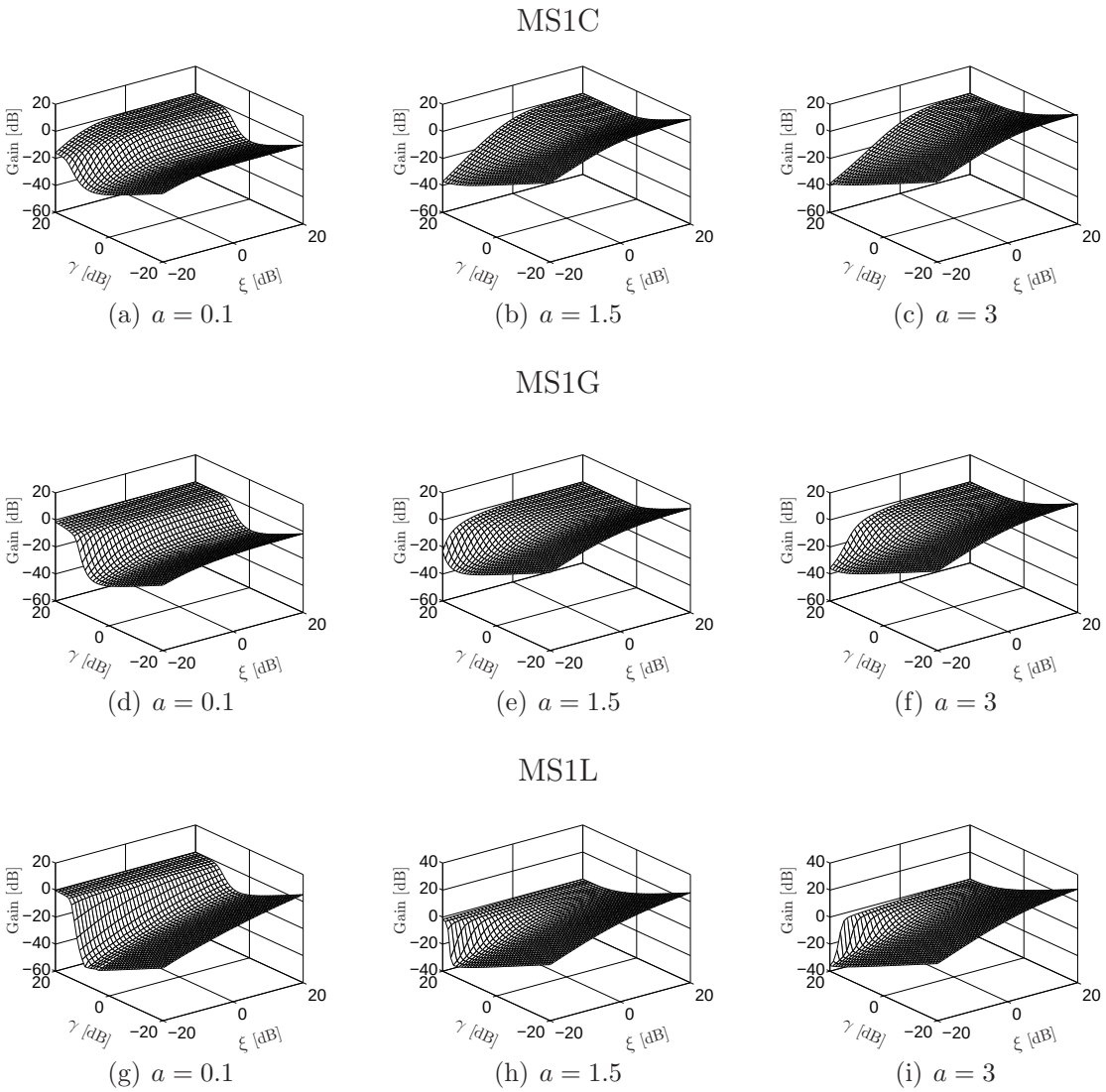
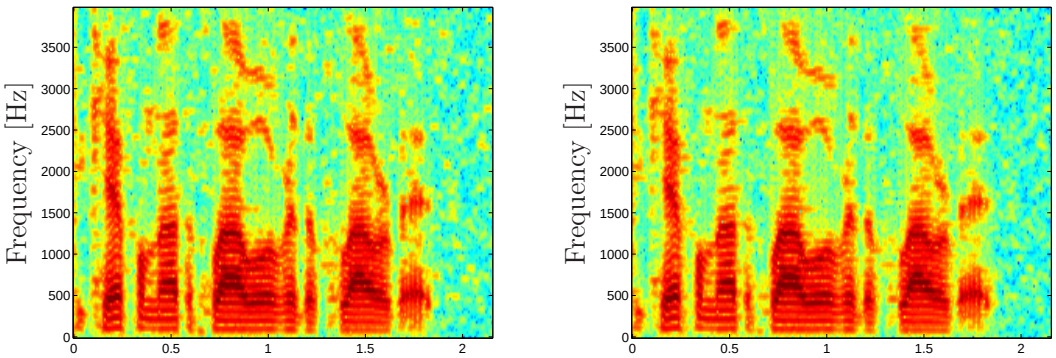


Figure 5.12: Suppression curves of the amplitude MMSE algorithms for different values of a as a function of the a priori SNR ξ and a posteriori SNR γ.

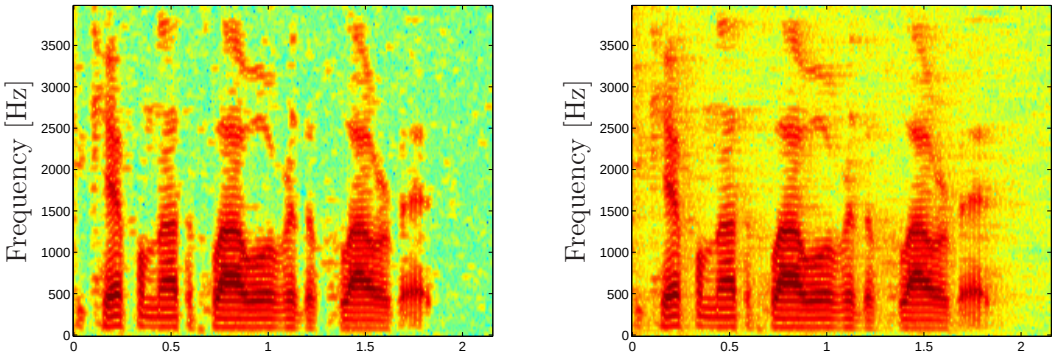
demonstrate the above behaviour are shown in figure 5.14.

Comparison w.r.t. the prior. The differences in the quality of the speech enhanced with the different priors, when the values of the shape parameter a result in equal levels of residual noise, are relatively subtle. The use of a longer tailed prior (e.g. Lognormal) results in a slightly better restoration of some weaker speech spectral components, especially at the onset of speech. A shorter tailed prior, such as the Chi on the other hand, results in smoother spectral peaks in the noise dominated regions of the spectrogram, and hence, the residual noise of the enhanced sentence is more uniform.



(a) a=0.2 (b) a=3

Figure 5.13: Speech enhanced with the MS2G algorithm for 2 different values of a. Increasing a reduces the intensity of the musical noise spectral peaks, but some of the speech spectral components are also suppressed.



(a) a=0.2 (b) a=3

Figure 5.14: Speech enhanced with the MS1C algorithm for 2 different values of a. Increasing a elevates the level of the residual noise, which eventually becomes uniform.

Comparison between the MAP and the MMSE estimators A comparison between the MAP and the amplitude MMSE estimators reveals that the MAP estimators result clearly in lower levels of residual noise. However, the preservation of the speech spectral components is better when the MMSE estimator is used and the resulting speech sounds less bandlimited and more natural. We should also note at this point that the computational complexity of the MMSE algorithms is generally higher compared to that of their MAP counterparts, because the former involve the calculation of special functions or numerical integration techniques.

5.3.3 Conclusion

Among the elements of the algorithms presented (STFT feature, estimator, prior) the one with the greatest impact on the performance was the estimator. Using values of a from the second range, the MAP estimator resulted in lower levels of residual noise compared to the MMSE estimators of the amplitude, while the latter were more successful in preserving the speech spectral components. The DFT MMSE estimators on the other hand, failed to result in uniform residual noise for any value of a, which could pose a significant problem in their employment in audio speech enhancement applications.

The selection of the STFT feature had a small impact on the algorithms that use the MAP estimator. The similarities between the expressions for the MP2C (eq. 3.11) and the MP1C (eq. 3.26) algorithms and between the MP2G (eq. 3.17) and the MP1G (eq. 3.31) algorithms also support this observation. The amplitude MAP algorithms however, were marginally better in the preservation of speech, so they might be preferred over their DFT counterparts. On the contrary, the selection of the STFT feature played an important part when the MMSE estimator was used, leading to musical residual noise and inferior speech restoration when the Re and Im parts were estimated instead of the amplitude. An advantage of the amplitude domain algorithms is that half the data load needs to be processed compared to their DFT counterparts.

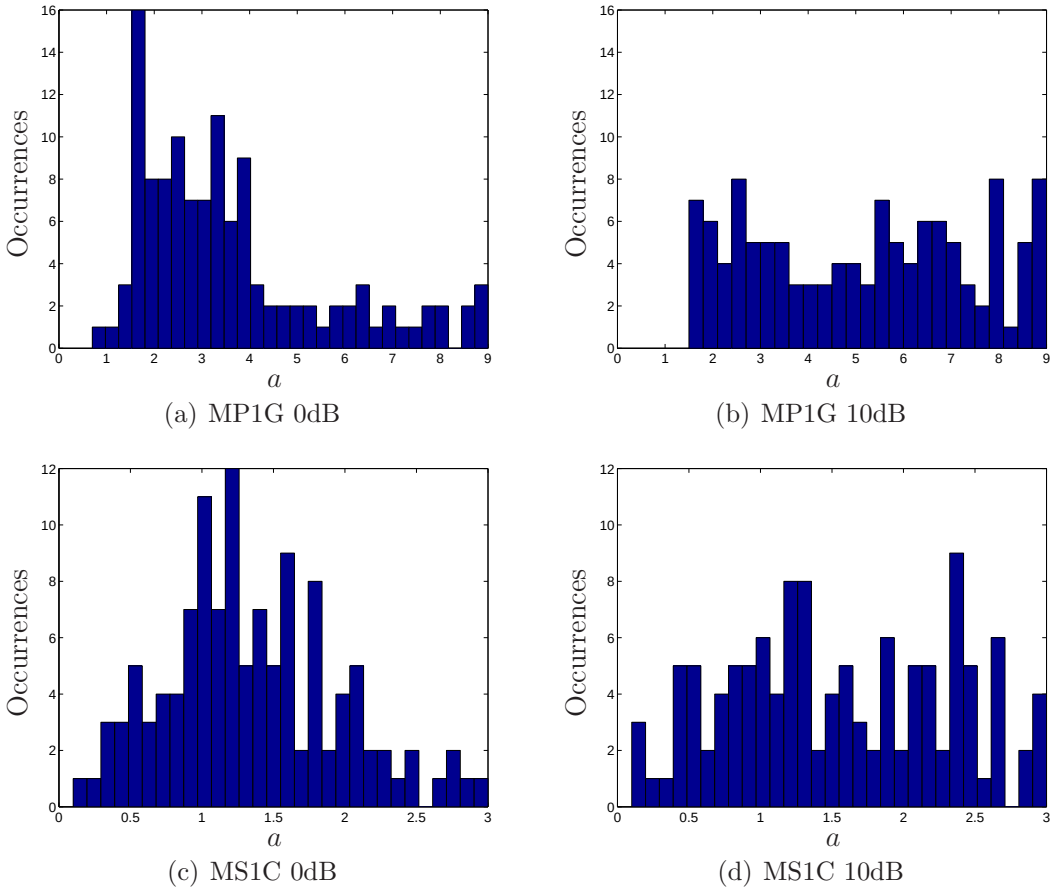
The choice of the prior had a rather moderate effect on the algorithms. This should be attributed to the fact that the flexibility provided by tuning the shape parameter a, offered the possibility of matching closely the performances achieved with the different priors. Combined with the MAP estimator, the Gamma priors achieved a good balance between the preservation of speech spectral components and the suppression of the spurious spectral peaks that are perceived as speech distortions. When the MMSE estimator was used, the Chi priors offered the most uniform background noise for an almost identical preservation of speech and residual noise level.

5.4 Subjective estimation of an optimal value for a

In the previous section we saw that among the three features of the examined algorithms (estimator, STFT feature and prior) the most influential in the performance of the algorithms was the estimator, while the estimated feature and the different prior densities had a somewhat less important role. Another critical component of the presented speech enhancement algorithms is the value of the shape parameter a. The analysis of the previous section showed that the value of a essentially determines the trade off between the musical character and the level of the residual noise and, to some extent, the preservation of the weaker speech spectral components.

In this section we present the results from a formal subjective listening test that we carried out, in order to identify a set of values for the shape parameter that result in the highest quality of speech. A set of 20 subjects were asked to determine the value of a that provided the best enhanced speech quality, using 6 sentences that were corrupted with white Gaussian noise at 0 and 10 dB input SegSNR. The subjects were presented with the clean and the noisy speech, and could then adjust the value of the shape parameter and listen to the corresponding enhanced speech. No visual cues were given for the value of a, so the subjects had to base their decision solely on the audio samples. Additionally, the value of a was randomised for each new sentence so that the preferred values for one sentence could not affect the decision made for the others. Finally, each subject was presented with a sequence of audio samples in which the order of the sentences and the input SegSNR levels was random.

Informal listening tests in §5.3 revealed that the MAP algorithms produced speech of very similar quality when the shape parameter of the priors was tuned so that the level of the residual noise was equal among the different MAP algorithms. For this reason, and in order to reduce the duration of the subjective experiment one MAP algorithm was evaluated, the MP1G. The latter was selected because it is computationally more efficient than the DFT MAP algorithms and results in a good trade off between the uniform character of the residual noise and the preservation of speech. From the MMSE family of algorithms we considered the estimators of the amplitude only, because of their tendency to generate uniform residual noise. The MMSE amplitude algorithm we selected was the MS1C, because it results in



(a) MP1G 0dB (b) MP1G 10dB

(c) MS1C 0dB (d) MS1C 10dB

Figure 5.15: Histograms of values of a selected in the subjective experiment.

the most uniform noise among the other MMSE amplitude algorithms, for a given residual noise level and speech restoration quality, while it is also the most efficient computationally.

In figure 5.15 we present the results of the subjective test for the two different algorithms and the two different noise levels. The histograms show the occurrences of the different values of a for all the subjects and the presented sentences. It is noticeable that for the low input SegSNR the selections are concentrated around some particular values. For the MAP algorithm the majority of the values are around 2-4, while for the MMSE the most popular range is between 1 and 1.5. For the higher input SegSNR on the other hand, the selected values are considerably more spread.

To verify the validity of the above observation we performed a chi square significance test [75]. The above test was used in order to check the hypothesis that the

Input SegSNR	0 dB		10 dB	
Algorithm	$\mathcal{X}^2$	p-value	$\mathcal{X}^2$	p-value
MP1G	103.0	1.5×10^{-13}	16.3	0.63
MS1C	68.0	2×10^{-7}	23.7	0.17

Table 5.1: Chi square significance test results for the two algorithms and input SegSNR levels.

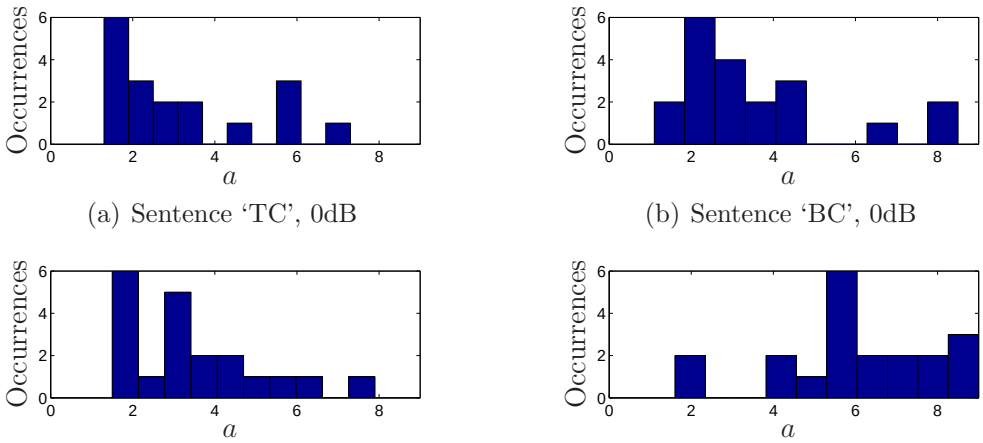
subjective test data for each algorithm and input SegSNR level came from a uniform distribution (null hypothesis). Rejection of the null hypothesis for the low SegSNR data and failure of rejection for the high SegSNR data would indeed confirm the larger spread of the high SegSNR data set.

Table 5.1 shows the results of the chi square test. The chi square statistic $\mathcal{X}^2$ is given by the formula

$$\mathcal{X}^2 = \sum_{i=1}^{N_\mathcal{X}} \frac{(O_i - \bar{O}_i)^2}{\bar{O}_i} \tag{5.2}$$

where O_i is the number of occurrences in the i^{th} histogram bin and $\bar{O}_i$ is the number of expected occurrences in the i^{th} histogram bin according to the assumed distribution (uniform). The number of histogram bins was $N_\mathcal{X} = 20$, which satisfies the requirement for a minimum of 5 expected occurrences in each of the histogram bins, given that the total number of observations for each case was 120. The p-value of the test denotes the probability of a random variable that follows the assumed distribution to have a chi square statistic larger than the respective statistic of the data. The smaller the p-value therefore, the stronger the evidence is for the rejection of the null hypothesis. The p-values shown in table 5.1 indicate that the null hypothesis can be safely rejected for the low input SegSNR condition for both algorithms, while for the high input SegSNR level there is not sufficient evidence for its rejection.

The differences in the shapes of the distributions for the two input SegSNR conditions can be attributed mainly to two reasons: The first is related to the fact that for the low input SegSNR condition the extreme values of a were not favoured, because either the musical noise was too intense (small a) or the residual noise was excessive (large a). For the high input SegSNR however, the effect of selecting a value of a closer to the extremes of the range was not as adverse, which generally



(a) Sentence 'TC', 0dB

(b) Sentence 'BC', 0dB

(c) Sentence 'TC', 10dB

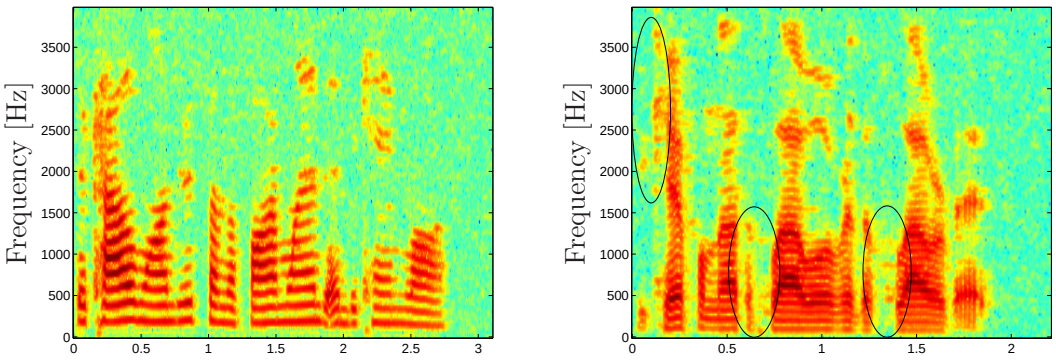
(d) Sentence 'BC', 10dB

Figure 5.16: Histograms of values of a selected in the subjective experiment for two different sentences and input SegSNR levels.

made harder to pinpoint an optimal value for a and contributed to the flatter shape of the respective histograms.

The second reason was related to the spectral content of some particular sentences. Specifically, it was observed that for two out of six sentences the subjects consistently chose higher values of a for the high input SegSNR compared to their selections for the low input SegSNR condition. An example is shown in figure 5.16 where the histograms of the selected values for two sentences and two input SegSNR's are shown. The results correspond to the MP1G algorithm, while the first sentence is *'The cow wandered from the farmland and became lost'*, denoted as 'TC' and the second is *'Be careful not to plough over the flower beds'* denoted as 'BC'. Note that for both input SegSNR levels the values selected for the first sentence are relatively similar (figures 5.16(a), 5.16(c)), while for the second sentence (figures 5.16(b), 5.16(d)) the values chosen for the high input SegSNR condition were significantly higher than those selected for the low SegSNR.

A retrospective evaluation of the above sentences, in terms of inspecting the respective spectrograms and performing informal listening tests, revealed that the above observations may stem from the differences in the distribution of the spectral energy of each sentence on the time frequency plane. For example, processing the 'BC' sentence with the MP1G algorithm and $a = 4$ at 10 dB input SegSNR resulted in a number of spurious spectral peaks, which were the result of the distribution of



(a) Sentence 'TC' (b) Sentence 'BC'

Figure 5.17: Spectrograms of two different sentences enhanced with the MP1G algorithm and $a = 4$. The ellipses highlight the spurious spectral peaks that are perceived as musical noise.

the weaker speech spectral components in the clean sentence. The spurious spectral peaks of the 'TC' sentence, when processed with the same algorithm and value of a, were considerably less. The spectrograms of the two sentences are shown in figure 5.17, where the spurious spectral peaks of the 'BC' that are perceived as musical noise are highlighted. The existence of more randomly placed spectral peaks in the 'BC' sentence, may have led the subjects to increase the level of the residual noise, by means of increasing the value of a for masking purposes. For the low input SegSNR the majority of the weak speech spectral components were immersed in noise and their recovery was not possible, while any remaining spurious spectral peaks were masked from noise for much smaller values of a due to the higher background noise level. This resulted in more consistent choices of a for the different sentences in the low SegSNR condition and consequently, the values of a are more concentrated in the respective aggregate histograms of figure 5.15.

Based on the results of the subjective test we propose as 'optimal' the mean values of the parameter a extracted from the low SegSNR condition[5]. These values were 2.6 for the MP1G and 1.4 for the MS1C algorithm. In table 5.2 we show the results in the objective measures for the two examined algorithms with the aforementioned values of a. Additionally, table 5.2 shows the results for the remaining MAP and amplitude MMSE algorithms for values of a that result in the same residual noise

[5]For the MAP algorithms we actually used the mean value of the data that appear in the main 'hump' between 1 and 4.5.

Input SegSNR		0 dB		10 dB		20 dB	
		SegSNR	PESQ	SegSNR	PESQ	SegSNR	PESQ
MP1C	$a = 1.6$	7.03	2.79	12.70	3.27	20.09	3.79
MP1G	$a = 2.6$	7.19	2.74	12.96	3.24	20.40	3.78
MP1L	$a = 1.1$	7.35	2.68	13.16	3.22	20.63	3.79
MP2C	$a = 1.1$	6.72	2.75	12.23	3.22	19.49	3.76
MP2G	$a = 2.0$	6.81	2.68	12.49	3.20	19.85	3.74
MS1C	$a = 1.4$	6.71	2.81	13.03	3.33	20.69	3.82
MS1G	$a = 2.4$	6.96	2.82	13.38	3.34	21.09	3.82
MS1L	$a = 1.4$	7.09	2.84	13.57	3.35	21.27	3.81
MS2C	$a = 2.0$	6.22	2.68	11.52	3.09	18.66	3.65
MS2G	$a = 2.5$	6.93	2.70	12.53	3.14	19.90	3.72

Table 5.2: Results in the objective measures obtained with values of a selected from the subjective test. The corrupting noise was white Gaussian.

Input SegSNR		0 dB		10 dB		20 dB	
		SegSNR	PESQ	SegSNR	PESQ	SegSNR	PESQ
MP1C	$a = 1.6$	10.52	3.37	16.63	3.81	23.88	4.18
MP1G	$a = 2.6$	10.62	3.36	16.79	3.80	24.02	4.18
MP1L	$a = 1.1$	10.66	3.36	16.84	3.82	24.03	4.18
MP2C	$a = 1.1$	10.20	3.34	16.25	3.79	23.44	4.17
MP2G	$a = 2$	10.20	3.32	16.38	3.78	23.58	4.16
MS1C	$a = 1.4$	9.72	3.37	16.61	3.82	24.13	4.20
MS1G	$a = 2.4$	9.91	3.38	16.86	3.82	24.39	4.20
MS1L	$a = 1.4$	9.99	3.38	16.95	3.81	24.49	4.18
MS2C	$a = 2$	9.73	3.23	15.69	3.70	22.83	4.12
MS2G	$a = 2.5$	10.45	3.29	16.48	3.76	23.71	4.15

Table 5.3: Results in the objective measures obtained with values of a selected from the subjective test. The speech was corrupted with car noise.

levels as the MP1G and MS1C algorithms respectively. Such a normalisation was not possible for the DFT MMSE algorithms, because the level of the residual noise remains almost constant for different values of a, as we mentioned in §5.3.2. For this reason in table 5.2 we present the results for the MS2C and MS2G algorithms with values of a empirically chosen so that they provide an adequate trade off between the musical character of the residual noise and the suppression of the speech spectral components. Table 5.3 shows the respective results for the car noise.

The objective scores reveal that the amplitude algorithms achieve higher results compared to their DFT counterparts. The same is also true for the MMSE algorithms compared to the MAP, with the exception of the SegSNR measure for the 0 dB input SegSNR level. The reason is that the MAP algorithms achieve a lower residual noise level, at the expense of the suppression of some spectral components that belong to speech. Regarding the priors we can note that the longer tailed priors (i.e. Lognormal) achieve higher SegSNR scores, mainly due to the better preservation of speech especially at its onset. Although the PESQ scores of the MMSE algorithms are fairly similar for the different priors, for the MAP algorithms higher PESQ scores are achieved for the short tailed priors (i.e. Chi). We believe that the reason is the preservation of some weak spectral components with the shorter tailed priors, which however contribute more to the musical character of the residual noise rather than to the enhancement of the speech quality.

5.5 Results for adaptively estimated values of a

In the previous sections the algorithms employed a fixed value of the parameter a for the entire duration of the speech utterances. In this section we evaluate the adaptive scheme for the estimation of a, which was presented in §4.4.

The values of a were estimated from a function of the kurtosis of the clean speech spectral samples, which for the DFT algorithms was defined as

$$\kappa_2 = \frac{\mathrm{E}[X^4] - 6\mathrm{E}[S^2]\mathrm{E}[N^2] - 3\left(\mathrm{E}[N^2]\right)^2}{\left(\mathrm{E}[S^2]\right)^2} \tag{5.3}$$

and for the amplitude algorithms as

$$\kappa_1 = \frac{\mathrm{E}[R^4] - 4\mathrm{E}[A^2]\mathrm{E}[B^2] - 2\left(\mathrm{E}[B^2]\right)^2}{\left(\mathrm{E}[A^2]\right)^2} \tag{5.4}$$

We will now discuss the method for estimating the moments, which are involved in the above expressions. The noise moments $\mathrm{E}[N^2]$ and $\mathrm{E}[B^2]$ can be estimated directly from the noise estimation algorithm. For example, if an estimate of the noise variance is $\widehat{\mathrm{E}[|\mathbf{N}|^2]}$, then we can set $\mathrm{E}[N^2] = \widehat{\mathrm{E}[|\mathbf{N}|^2]}/2$ and $\mathrm{E}[B^2] = \widehat{\mathrm{E}[|\mathbf{N}|^2]}$.

For the estimation of the fourth moment of the noisy speech coefficients we used a first order recursive averaging. If we define an estimator $\widehat{\mathrm{E}[X^4]}$ for the fourth moment of the noisy speech DFT coefficients $\mathrm{E}[X^4]$, an estimate can be obtained as

$$\widehat{\mathrm{E}[X^4]}_{|k} = (1-\lambda)\widehat{\mathrm{E}[X^4]}_{|k-1} + \lambda X^4 \tag{5.5}$$

where the subscripts k and $k-1$ indicate the current and previous time frames respectively. The fourth moment estimator for the noisy speech amplitude coefficients $\mathrm{E}[R^4]$ was defined accordingly. The smoothing parameter λ was found to have a major influence on the performance of the adaptive scheme. Large values (~ 0.1) resulted in highly fluctuating estimates of a. The application of these values to the algorithms resulted in speech that suffered from high levels of background noise with a strong musical character. Too small values of a (~ 0.0001) resulted in an insensitivity of the adaptive scheme to the speech changes. The values of λ that were found to give the optimal results were in the range of 0.001 to 0.01. In all the simulations the value used was $\lambda = 0.005$.

The second moments of the speech samples were obtained from a smoothed version of the a priori SNR. When unsmoothed values of the a priori SNR were used, the resulting speech suffered from high levels of musical background noise. The smoothing was performed with a recursive averaging estimator, in a similar fashion to eq. 5.5. The estimators for $\mathrm{E}[S^2]$ and $\mathrm{E}[A^2]$ were

$$\widehat{\mathrm{E}[S^2]}_{|k} = (1-\lambda)\widehat{\mathrm{E}[S^2]}_{|k-1} + \lambda\,\xi\,\mathrm{E}[N^2] \tag{5.6}$$

and

$$\widehat{\mathrm{E}[A^2]}_{|k} = (1-\lambda)\widehat{\mathrm{E}[A^2]}_{|k-1} + \lambda\,\xi\,\mathrm{E}[B^2] \tag{5.7}$$

respectively. The same parameter λ as in eq. 5.5 was used for the smoothing.

The estimates of a were permitted to take values only within a certain range. For the Chi and Gamma priors this range was [0.01, 3] and for the Lognormal priors the range was [0.1, 3]. The value of the lower limit was not particularly important, as the estimates rarely were below that. However, when lower estimates were allowed, their influence was found to be rather damaging, as it resulted in the excessive suppression of some speech spectral components. The value of the upper limit played a more important role. Firstly, as it was indicated by informal listening tests, values of a beyond the above limits were not found to improve the speech restoration in some way. Additionally, the adaptive scheme occasionally produced estimates of a that had unusually high values. These were mostly due to poor estimation because of high background noise levels. For these values of a the algorithms resulted in excessively high background noise levels and bounding the values of a mitigated the above problem to some extent. Furthermore, very large values of a caused numerical issues with the routines that calculated the special functions or with those that performed the numerical integrations. Bounding the values of a alleviated this problem as well.

Tables 5.4, 5.6 show the results in the objective measures obtained with the MAP and MMSE algorithms respectively, using the adaptive method for the estimation of a. The results are presented in the column under the header 'adaptive'. To provide a comparison with the case when a fixed value of a is used, we used the algorithms with a fixed a, which was equal to the median of the values of a estimated with the adaptive scheme, across all the time and frequency samples. The results are shown in the columns under the header 'fixed'. The respective median values for the MAP and MMSE algorithms respectively are shown in tables 5.5, 5.7. A comparison between the 'adaptive' and 'fixed' scores shows that the results are not drastically different, although the scores obtained with a fixed value of a tend to be somewhat higher.

Figure 5.18 shows the behaviour of the estimates of a in a single frequency bin. The typical behaviour of a is that it increases at the onset of speech and then drops due

	White Noise				Car Noise			
	SegSNR		PESQ		SegSNR		PESQ	
	Fixed	Adaptive	Fixed	Adaptive	Fixed	Adaptive	Fixed	Adaptive
	0 dB Input SegSNR							
MP1C	7.40	7.20	2.71	2.67	11.05	10.24	3.44	3.44
MP2C	7.15	6.96	2.71	2.67	10.82	9.96	3.42	3.42
MP1G	7.60	7.04	2.69	2.67	11.12	10.34	3.41	3.44
MP2G	7.50	7.18	2.68	2.63	10.90	10.23	3.37	3.36
MP1L	7.41	7.40	2.62	2.63	10.80	10.58	3.35	3.35
	10 dB Input SegSNR							
MP1C	13.75	13.64	3.38	3.37	17.21	16.85	3.93	3.94
MP2C	13.58	13.45	3.38	3.37	17.00	16.67	3.92	3.93
MP1G	13.68	13.71	3.33	3.38	17.15	16.94	3.91	3.94
MP2G	13.35	13.30	3.27	3.27	16.87	16.68	3.88	3.88
MP1L	13.20	13.18	3.19	3.17	16.84	16.76	3.84	3.83
	20 dB Input SegSNR							
MP1C	21.18	21.15	3.94	3.94	24.50	24.10	4.25	4.24
MP2C	21.01	20.93	3.95	3.94	24.24	23.90	4.24	4.24
MP1G	21.07	21.35	3.91	3.95	24.37	24.27	4.24	4.24
MP2G	20.63	20.68	3.87	3.87	24.03	23.82	4.23	4.22
MP1L	20.63	20.63	3.80	3.80	23.99	23.97	4.20	4.19

Table 5.4: Results with adaptively estimated values of a for the MAP algorithms.

Noise Type	White			Car		
SegSNR	0	10	20	0	10	20
MP1C	0.06	0.08	0.10	0.10	0.10	0.11
MP2C	0.04	0.05	0.07	0.07	0.07	0.07
MP1G	0.11	0.14	0.20	0.20	0.20	0.20
MP2G	0.09	0.14	0.18	0.19	0.19	0.19
MP1L	0.48	0.59	0.65	0.66	0.66	0.67

Table 5.5: Median of the values of a estimated with the adaptive scheme for the MAP algorithms.

to the recursive estimation of the speech moments. Informal listening tests indicate that the adaptive scheme results in speech which suffers from musical residual noise. The quality of the resulting speech is similar to that obtained using the fixed values of a shown in tables 5.5 and 5.7.

An additional drawback that is associated with the adaptive estimation scheme is that in the presence of excessive noise levels, the parameter a can take high

	White Noise				Car Noise			
	SegSNR		PESQ		SegSNR		PESQ	
	Fixed	Adaptive	Fixed	Adaptive	Fixed	Adaptive	Fixed	Adaptive
	0 dB Input SegSNR							
MS1C	7.76	7.75	2.85	2.86	10.97	10.50	3.53	3.53
MS2C	7.40	7.55	2.79	2.82	10.90	10.92	3.46	3.46
MS1G	7.79	7.78	2.86	2.86	10.98	10.77	3.53	3.51
MS2G	7.54	7.64	2.81	2.82	11.03	11.05	3.46	3.46
MS1L	7.70	7.60	2.88	2.87	10.68	10.35	3.49	3.46
	10 dB Input SegSNR							
MS1C	13.83	13.79	3.49	3.49	17.28	17.09	3.97	3.96
MS2C	13.45	13.42	3.41	3.40	16.98	16.94	3.93	3.93
MS1G	13.88	13.85	3.49	3.47	17.33	17.23	3.97	3.95
MS2G	13.51	13.52	3.40	3.39	17.11	17.07	3.93	3.92
MS1L	13.91	13.86	3.45	3.42	17.29	17.16	3.91	3.89
	20 dB Input SegSNR							
MS1C	21.24	21.20	3.99	3.98	24.48	24.45	4.26	4.26
MS2C	20.78	20.73	3.94	3.93	24.14	24.10	4.25	4.24
MS1G	21.32	21.32	3.99	3.98	24.55	24.58	4.26	4.25
MS2G	20.90	20.88	3.93	3.92	24.29	24.26	4.24	4.24
MS1L	21.43	21.41	3.94	3.93	24.65	24.60	4.23	4.22

Table 5.6: Results with adaptively estimated values of a for the MMSE algorithms.

Noise Type	White			Car		
SegSNR	0	10	20	0	10	20
MP1C	0.04	0.08	0.10	0.10	0.10	0.10
MP2C	0.02	0.05	0.06	0.07	0.07	0.07
MP1G	0.12	0.14	0.19	0.19	0.20	0.20
MP2G	0.07	0.14	0.18	0.19	0.19	0.19
MP1L	0.51	0.61	0.66	0.66	0.66	0.67

Table 5.7: Median of the values of a estimated with the adaptive scheme for the MMSE algorithms.

values in some frequency bins, which results in high levels of residual noise at those frequencies. An example is shown in figure 5.19, where the spectrogram of the enhanced sentence *Cyclical programs will never compile* and the estimated values of a are shown. The above sentence was corrupted with white Gaussian noise at 0 dB input SegSNR and enhanced with the MP1G algorithm. Note that in some of the high frequencies the estimation scheme returns large values of a, which result in high levels of bandlimited noise.

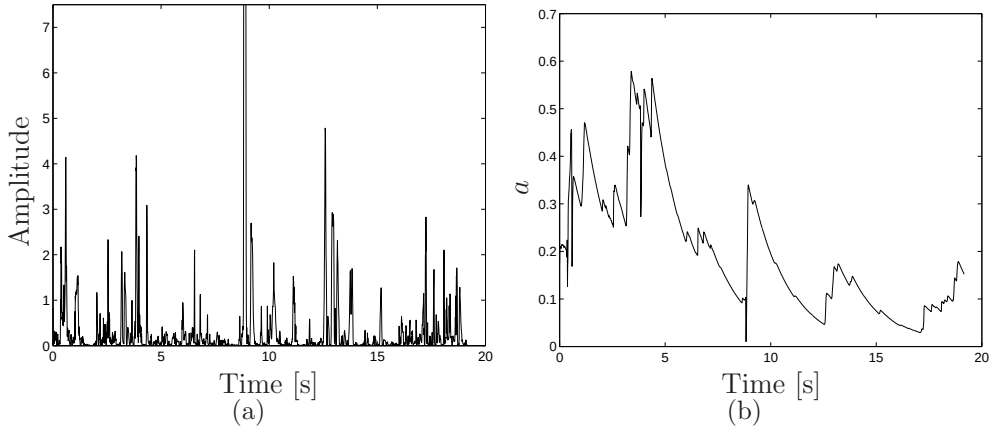


Figure 5.18: (a) Clean speech spectral amplitude values from the frequency bin corresponding to 1 KHz, (b) Values of a estimated with the adaptive scheme from the corresponding noisy speech data and the MP1G algorithm.

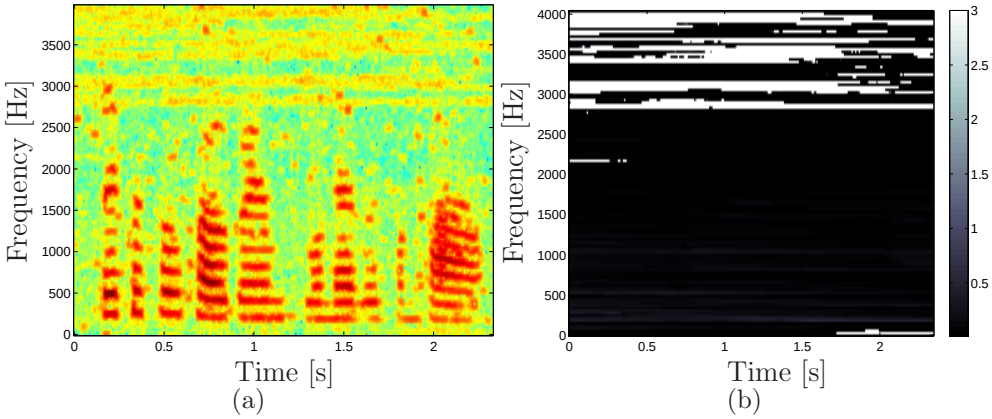


Figure 5.19: (a) Spectrogram of speech enhanced with the MP1G algorithm and the adaptive scheme (b) Estimated values of a with the adaptive scheme.

Overall, the scores in the objective measures of the speech processed with the adaptive scheme were among the highest that could be obtained using fixed values of a. Nevertheless, the residual noise had a rather strong musical character, while poor estimation occasionally resulted in high levels of narrowband residual noise.

5.6 Summary

In this chapter we evaluated the algorithms that constitute the framework of algorithms that was proposed in chapter 3. The performance of the algorithms was evaluated using the SegSNR and PESQ objective measures, as well as with formal and informal subjective listening tests. All the algorithms employed the DD method for the adaptive estimation of the speech priors' scale parameter θ, because fixed values of θ were shown to result in high levels of musical noise. As the method for the estimation of the shape parameter a, which was described in §4.3.3 and is compatible with the DD method failed to result in consistent estimates, the performance of the algorithms was measured as a function of the value of the shape parameter a. The adaptive method for the estimation of a, which was proposed in §4.4 was also evaluated in this chapter.

The most influential element of the proposed speech enhancement algorithms was shown to be the estimator. On the other hand, the estimated feature or the employed prior had a somewhat smaller influence on the results. The impact of the different features of the algorithms in their performance was summarised in §5.3.3.

The priors' shape parameter a essentially controlled a trade off between the musical character of the residual noise and its level, while the preservation of the weaker speech spectral components was influenced to some extent. Small values of a resulted in an adequate preservation of speech, but the residual noise had a strong musical character. Larger values of a resulted in a more uniform residual noise, but generally led to an increase in its level. The increase in the values of a also deteriorated the preservation of the weaker speech spectral components, especially for the MAP algorithms. Optimal values of a were sought by means of a formal subjective listening test.

The scheme for the adaptive estimation of a resulted in some of the highest scores

in the objective measures that could be achieved with any fixed value of a. The speech however suffered from musical residual noise, while occasional poor estimates of a, mainly due to low input SegSNR levels, resulted in poor suppression of the background noise in some relatively narrow frequency bands.

Chapter 6

Noise estimation

The noise estimation algorithm is a vital part of an integrated speech enhancement scheme. Accurate noise estimates are critical in such a scheme's overall performance, because an underestimation can result in less suppression of the background noise, while noise overestimates result in increased speech distortion. In this chapter we review some of the most well known noise estimation algorithms and present two novel noise estimation algorithms that have been developed as part of this project.

In chapter 3 we examined a number of different prior densities that were used to model the speech data. The noise STFT coefficients on the other hand, were modelled only with a single distribution, the complex Gaussian. If the background noise is stationary the Gaussian assumption for the distribution of the noise STFT coefficients is supported by the Central Limit Theorem [15], and can be easily verified by simulations. The above assumption is also valid even if the noise is only approximately stationary. However, if the noise is time varying and its amplitude fluctuates significantly with time, the distribution of its STFT coefficients can deviate significantly from the Gaussian model.

In the early stages of this project we experimented with alternative noise models and we used the Gaussian Mixtures Models (GMM's) that fitted the time varying noise data very accurately. When these models were combined with a speech enhancement algorithm they resulted in some improvement over the models that used a single Gaussian distribution with a fixed variance. However, when the single Gaussian models were allowed to adapt their variance according to the changes in the power

of the noise signal their performance surpassed that of the GMM's. Results of this work are presented in §6.2.

The above results showed that the Gaussian model can be adequate for time varying noise[1] as well, as long as there is an algorithm that not only can estimate the noise power, but it can also track its changes with time. Such an algorithm has been developed during the course of this project and is presented in §6.3. The proposed algorithm exploits some similarities between the distribution of the noisy speech spectral amplitude coefficients within a single frequency bin and the distribution of the corresponding coefficients of the corrupting noise. The above similarities are used for the extraction of samples, from a window of past spectral amplitudes of noisy speech, which are more likely to contain noise only. These samples are then used for the calculation of an estimate of the noise power. The extraction of the noise samples is based on matching the two first moments of the Rayleigh distribution.

The algorithms that are presented in this chapter are all designed to estimate the power of the amplitude of the noise STFT coefficients i.e. $\mathrm{E}[|\mathbf{N}|^2]$, or $2\sigma_N^2$ in the notation of chapter 3. For use with algorithms that enhance the Re and Im parts of the noisy STFT coefficients the output of the noise estimation algorithms has to be divided by two. In order to simplify the notation, in this chapter we introduce the symbol $\mathcal{P}$, which will denote the estimate of $\mathrm{E}[|\mathbf{N}|^2]$, that is $\mathcal{P} \equiv \widehat{\mathrm{E}[|\mathbf{N}|^2]}$. We begin our discussion on the estimation of noise by summarising in the next section some of the most prominent noise estimation methods that can be found in the literature, highlighting also their main benefits and drawbacks.

6.1 Previous methods

6.1.1 Noise estimation by averaging past spectral values

One method for estimating the noise power is found by averaging the past values of the noisy speech spectrum. The averaging, which is typically implemented with a first order recursion, is performed with the samples that are assumed to belong

[1] We need to mention here that the noise should vary with time in such a way that the signal is stationary over the duration of an STFT analysis window. These signals might also be called quasi stationary.

to noise only, while the noise estimate is not updated when speech is believed to be present. The above rule can be written as

$$\begin{aligned} H_0(k,l) : \mathcal{P}(k,l) &= \alpha_d \mathcal{P}(k,l-1) + (1-\alpha_d) R(k,l)^2 \\ H_1(k,l) : \mathcal{P}(k,l) &= \mathcal{P}(k,l-1) \end{aligned} \tag{6.1}$$

$H_0(k,l)$ denotes the hypothesis that $R(k,l)$ contains noise only, while $H_1(k,l)$ denotes the hypothesis that $R(k,l)$ contains noisy speech. α_d is the smoothing parameter.

The speech presence or absence can be determined with a VAD. A simple implementation of a VAD is achieved by measuring the spectral distance between the frame that needs to be classified and a noise template, which has to be provided. If the distance is above a threshold the frame is assumed to contain speech, while it is considered to contain only noise otherwise. For example suppose we wish to classify frame l and we are provided with a noise template $\hat{B}(k,l)$, which can be the noise estimate from the previous frame. A decision rule for the speech presence or absence can be the following

$$\frac{1}{K} \sum_{k=1}^{K} \max \Big(20 \log(R(k,l)) - 20 \log(\hat{B}(k,l)), 0 \Big) \underset{H_0(l)}{\overset{H_1(l)}{\gtrless}} \delta \tag{6.2}$$

$H_1(l)$ denotes the speech presence in frame l and $H_0(l)$ the speech absence, while δ is an empirically determined threshold.

An improved version of a VAD was proposed by Sohn et al. [91]. A ratio of the likelihood of speech presence given the noisy measurements over the likelihood of speech absence is first formed and the average of the ratio's logarithm is then compared to a threshold. The rule can be written as follows

$$\frac{1}{K} \sum_{k=0}^{K-1} \ln \left(\frac{p(R(k,l)|H_1(k,l))}{p(R(k,l)|H_0(k,l))} \right) \underset{H_0(l)}{\overset{H_1(l)}{\gtrless}} \delta \tag{6.3}$$

where K is the number of the frequency bins. The likelihoods are formed with the assumption that the STFT coefficients of speech and noise have a Gaussian distribution, according to the analysis in [31].

A problem with estimating noise using VAD's is that if the noise level increases suddenly, it is possible that the spectral distance or the likelihood ratio for the noise only frames exceeds the threshold δ and noise only frames are misclassified as speech. Additionally, it may be possible that the number of frames that are classified as containing noise only is very small and an accurate noise estimate cannot be obtained.

Malah et al. [69] proposed a scheme that offers better adaptability to the changes of the noise power. Specifically, an adaptive smoothing factor α_d was proposed that is a linear function of the negative value of the a posteriori SNR γ. The rationale was that higher values of γ might indicate an increase in the noise power and therefore the value of α_d should decrease in order to achieve a quicker adaptation of the noise power estimate. Nevertheless, as the above algorithm is applied only in the speech absent frames, the use of a VAD is still required.

The method proposed by Hirsch and Ehrlicher [48] did not explicitly employ a VAD but there was an implicit decision about the presence of speech. The noise estimate was updated only when the noisy speech spectral amplitude $R(k, l)$ was below a threshold, which was directly proportional to the square root of the existing estimate of the noise power. It follows, that sudden increases in the level of the noise power would cause this method to stop updating the estimates of noise.

A conceptually similar method was proposed by Lin et al. [64], which did not use a VAD, in the sense that a hard decision about the presence or absence of speech was not required. The first of the equations in 6.1 was employed for updating the noise estimates, with a variable value of α_d. The value of α_d was a sigmoid function of the a posteriori SNR γ, which was zero for $\gamma < 1$ and 1 for $\gamma \gtrsim 2$. The behaviour of α_d under this estimation scheme is the opposite to that proposed in [69]. In the latter work, the value of α_d dropped with increasing γ because the update was performed only on the frames that were classified as noise from the VAD. The goal in that case was to track the increasing levels of noise, in frames that were already classified as noisy. On the other hand, in [64] the value of α_d increases with γ because it implements a soft decision alternative to the VAD.

6.1.2 Minimum statistics noise estimation

A popular method for estimating the power of the background noise is given by tracking the minima of the amplitude of the noisy STFT coefficients within a frequency bin. This family of methods is based on the observation that the noise power in a frequency bin is related to the minimum values of the STFT coefficients. Indeed, if we consider the amplitude values of a clean speech frequency bin, its minimum values should be found during speech pauses and should ideally be zero. When adding background noise, the minima increase and their values are related to the average noise power.

The first algorithm that made use of the minimum statistics estimation method was proposed by Martin [70]. To avoid problems related with the minima outliers, the noisy speech power spectrum was first smoothed with a first order recursive equation with a constant smoothing factor

$$\mathcal{R}(k,l) = \alpha_p \mathcal{R}(k,l-1) + (1-\alpha_p) R(k,l)^2 \tag{6.4}$$

where $\mathcal{R}(k,l)$ is the smoothed noisy speech power spectrum and α_p the smoothing factor. The minimum of $\mathcal{R}$ was then found in a window of $D = 100$ samples. As the values of the minima will necessarily be lower than the average of the noise power, the calculated minimum values were then compensated with a constant factor to yield an unbiased estimate.

Later, Martin [71] introduced some improvements to his original algorithm. The first was to introduce a variable smoothing factor α_p for the power spectral values. The reason was to avoid the compromise between insufficient smoothing of the samples that belonged to noise and widening the spectral peaks that belonged to speech. Additionally, he derived a variable compensation factor for the bias, using results from the theory of minimum statistics.

To improve the efficiency of the minimum searching algorithm, Martin also proposed to split the original window of D samples in D_s subwindows . In this way and by storing the minima of the $D_s - 1$ previous subwindows, the number of comparisons per signal frame, and thus the total computational cost, decreased significantly. Fur-

thermore, the search for local minima in the current subwindow was also proposed, in an effort to increase the speed of the algorithm's response during periods when the noise power is increasing.

An alternative method of searching for minima was proposed by Doblinger [29]. Instead of searching for minima within a window, the author proposed to track the minimum values as

$$\begin{aligned} \text{if} \quad \mathcal{R}_{\min}(k,l-1) &< \mathcal{R}(k,l) \\ \mathcal{R}_{\min}(k,l) &= \beta_1 \mathcal{R}_{\min}(k,l-1) + \frac{1-\beta_1}{1-\beta_2}\left(\mathcal{R}(k,l) - \beta_2 \mathcal{R}(k,l-1)\right) \\ \text{else} \\ \mathcal{R}_{\min}(k,l) &= \mathcal{R}(k,l) \end{aligned} \tag{6.5}$$

where $\mathcal{R}_{\min}(k,l)$ is the minimum value of the smoothed noisy speech power spectrum at the (k,l) time frequency point and β_1, β_2 are experimentally determined constants. The constant β_2 in particular, controlled the adaptation time of the minimum to changes in the noise power. This method of minimum tracking is reported in [83] to respond better in abrupt changes of the average noise power.

Two other methods that are related to the minimum statistics noise estimation are those proposed by Ris and Dupont [86] and by Stahl et al. [92]. The first method proposed to calculate the average of the $d < D$ lowest energy samples within a window of D samples and then compensate for the bias. The second method exploited the sample with the d^{th} lowest value (quantile) in the window of D samples.

6.1.3 Minima controlled recursive averaging noise estimation

Algorithms of this category estimate the noise variance through averaging past spectral values, in a similar sense to the algorithms presented in §6.1.1. However, the parameter that controls the averaging is determined by the minima of the power spectral values. Therefore, algorithms of this section, which also are the most recent, can be viewed as a hybrid between the algorithms of §6.1.2 and §6.1.1.

An algorithm of this category was presented by Rangachari et al. [84]. The noisy

speech spectrum is first smoothed according to eq. 6.4 with $\alpha_p = 0.7$. A comparison of $\mathcal{R}(k,l)$ with $\mathcal{P}(k,l-1)$ yields a rough decision about the presence of speech. If speech is judged to be absent (H_0) the noise is estimated with the first of eqs. 6.1 and a fixed $\alpha_d = 0.8$. If speech is present however, the parameter α_d is controlled by the minima. The minimum $\mathcal{R}_{\min}(k,l)$ is found with Doblinger's method and the ratio $\mathcal{R}(k,l)/\mathcal{R}_{\min}(k,l)$ is compared to a frequency dependent threshold. If the value of the ratio is below the threshold (higher probability of speech absence), the value of α_d remains at 0.8. Otherwise, it is more likely that the $(k,l)^{\text{th}}$ sample belongs to speech and α_d becomes 1, so that the noise estimate is not updated.

The algorithms by Cohen and Berdugo [23] and Rangachari and Loizou [83] take a slightly different approach. By introducing the conditional probability of speech presence $p(H_1^c(k,l)) \equiv p(H_1(k,l)|R(k,l))$, they modify the recursive averaging eqs. 6.1 as

$$\begin{aligned} \mathcal{P}(k,l) &= \mathcal{P}(k,l-1)p(H_1^c(k,l)) + \\ &\quad \left(\alpha_d \mathcal{P}(k,l-1) + (1-\alpha_d)R(k,l)^2\right)\left(1 - p(H_1^c(k,l))\right) \end{aligned} \tag{6.6}$$

In the above equation, the two branches of eq. 6.1 can be identified, multiplied with the conditional speech presence probability. As the traditional smoothing factor α_d is kept constant, we can see that the recursive equation is now controlled by $p(H_1^c(k,l))$.

The conditional speech presence probability is calculated as follows: The ratio $\mathcal{R}(k,l)/\mathcal{R}_{\min}(k,l)$ is compared to a threshold δ, and if it is found greater than the threshold, the indicator variable $Ind(k,l)$ takes the value 1, otherwise it becomes zero. The conditional speech presence probability is then calculated as

$$p(H_1^c(k,l)) = \alpha_{pr} p(H_1^c(k,l-1)) + (1-\alpha_{pr}) Ind(k,l) \tag{6.7}$$

where $\alpha_{pr} = 0.2$

A difference between [23] and [83] is that in the first the minima are estimated with Martin's method while in the second with Doblinger's. Additionally, the threshold δ in the second method is frequency dependent.

Finally, Cohen [19] proposed an alternative version of his previous algorithm by estimating the conditional speech presence probability within a Bayesian framework. Specifically, the speech presence probability given the noisy measurements was given by

$$p(H_1|\gamma) = \frac{p(\gamma|H_1)p(H_1)}{p(\gamma|H_1)p(H_1) + p(\gamma|H_0)p(H_0)} \tag{6.8}$$

where γ is the a posteriori SNR. The probabilities of γ given the speech presence or absence, were derived from the assumption of the Gaussian distribution of speech and noise STFT coefficients, following the model in [31]. The minima of the spectral values in this algorithm were used to control the probability of the speech absence $p(H_0)$.

6.1.4 Energy clustering noise estimation

The energy clustering noise estimation method is based on the analysis of histograms of the logarithm of the amplitude of consecutive STFT samples of noisy speech within a single frequency bin, and the corresponding histograms of the corrupting noise. Under the assumption that the analysed segments contain both speech and silence portions, the histograms can look like the ones depicted in figure 6.1. Observe that the distribution of noisy speech consists of two modes, the leftmost of which, corresponds to the samples that contain noise only, while the rightmost corresponds to the samples that contain speech plus noise. Additionally, the leftmost mode of the speech distribution is approximately at the same position with the mode of the distribution of the corrupting noise. Fitting two Gaussian pdf's with the EM algorithm [26], as proposed by Van Compernolle [95], or the utilisation of a two centroid algorithm such as the *k-means*, as proposed by Ris and Dupont [86], can extract the position of the leftmost mode of the noisy speech distribution, and hence, yield an estimate for the noise energy.

The main drawback of this algorithm is that the assumption of the existence of two modes in the distribution of noisy speech is not always valid, particularly when the input SNR is low. The merging of the two modes in these cases can result in gross inaccuracies in the noise estimates.

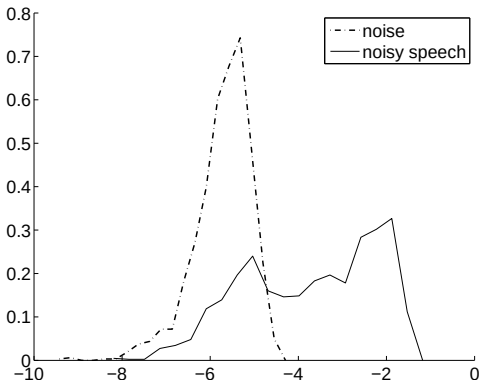
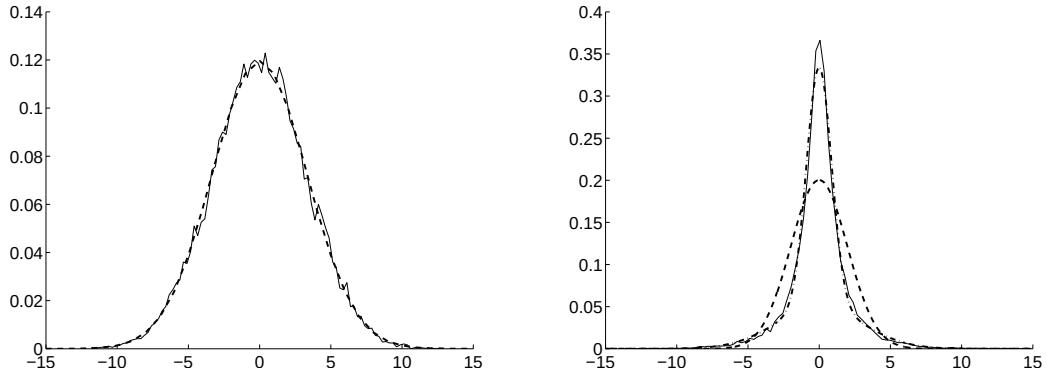


Figure 6.1: Histograms of the logarithm of the amplitude of noisy speech and noise. Samples were extracted from the frequency bin centered at 1 KHz from speech corrupted with white Gaussian noise at 10 dB SegSNR.

6.2 Noise estimation based on Gaussian Mixture Models.

According to Brillinger [15], the DFT samples of a stationary signal with finite moments follow a complex Gaussian distribution as the length of the DFT tends to infinity. The complex Gaussian distribution will have zero mean, while its variance at a particular frequency will be proportional to the power spectrum of the signal at the same frequency. According to the above argument, consecutive STFT samples within a single frequency bin of a stationary signal will be distributed according to approximately the same Gaussian distribution. Figure 6.2(a) shows the Re part of consecutive STFT samples of a stationary white Gaussian noise signal extracted from the same frequency bin, which was centered at 1 KHz. The fitting of a Gaussian distribution, with parameters estimated with the maximum likelihood method from the samples of the signal, is also shown. Indeed, the accurate fitting of the Gaussian distribution is in agreement with the theoretical result of Brillinger.

The Gaussian distribution however, is not an accurate model for consecutive STFT samples calculated from time varying noises. Figure 6.2(b) shows the histogram of the Re part of consecutive STFT samples calculated from time varying train noise and the maximum likelihood fit of a Gaussian distribution. The fitting is clearly poor. Based on the above observation and taking into account that the majority of the recorded noises exhibit at least some variation with time, we investigated the potential modelling of the noise STFT samples with Gaussian Mixture Models (GMM) [12] .



(a) (b)

Figure 6.2: (a) Histogram of stationary white Gaussian noise (continuous line) and ML fitting of a Gaussian distribution (dashed), (b) Histogram of time varying train noise (continuous line), ML fitting of a Gaussian distribution (dashed) and ML fitting of a 2 component GMM (dash-dot).

The Gaussian Mixture Models are density functions of the form

$$p(N) = \sum_{m=1}^{M} w_m \, \mathcal{N}(\mu_m, \sigma_m^2) \tag{6.9}$$

where $\mathcal{N}(\mu_m, \sigma_m^2)$ is a Gaussian density function with mean μ_m and variance σ_m^2. The weights w_m ensure that $\int p(N)\, dN = 1$ and M is the number of the Gaussian densities that constitute the mixture. The estimation of the parameters of the GMM is typically performed with the Estimation Maximisation (EM) algorithm [12, 26].

The model we proposed for the Re and Im parts of the noise STFT is a complex zero mean GMM of the form

$$p(N_{\mathrm{Re}}, N_{\mathrm{Im}}) = \sum_{m=1}^{M} \frac{w_m}{2\pi\sigma_{N,m}^2} \exp\left[-\frac{N_{\mathrm{Re}}^2 + N_{\mathrm{Im}}^2}{2\sigma_{N,m}^2}\right] \tag{6.10}$$

where N_{Re} and N_{Im} represent the Re and Im parts of the noise STFT coefficients. The above model implies that the Re and Im parts of noise have the same variance, which is in agreement with empirical observations. Integration of eq. 6.10 w.r.t. N_{Im} yields

$$p(N_{\mathrm{Re}}) = \sum_{m=1}^{M} \frac{w_m}{\sqrt{2\pi\sigma_{N,m}^2}} \exp\left[-\frac{N_{\mathrm{Re}}^2}{2\sigma_{N,m}^2}\right] \tag{6.11}$$

The distribution of the Im part is identical. The above equation reveals that the

proposed noise model does not assume that the Re and Im parts of noise are independent, as $p(N_{\mathrm{Re}}, N_{\mathrm{Im}}) \neq p(N_{\mathrm{Re}})p(N_{\mathrm{Im}})$. The dependencies reflect the fact that if the variance of, say, the Re part changes due to a change in the noise power, the variance of the Im part will change as well; hence the Re and Im parts are not independent.

The model parameters w_m and $\sigma^2_{N,m}$ are estimated separately from the Re and Im parts, based on equation 6.11 via the EM algorithm. Simulations show that the parameters estimated from the Re or Im parts are equal down to statistical fluctuations. Therefore, the estimates obtained from either the Re or the Im parts can be used with the model, or if both estimates are available their mean can also be used. In figure 6.2(b) the dash-dot line shows the fit of the GMM with $M = 2$ to the real part of the time varying train noise data.

In the following, we derive the MMSE estimator of the speech spectral amplitude with the 1 sided Chi priors and the proposed GMM noise model. Using eq. 6.10 and following the same procedure as in appendix A.1 we can show that the likelihood $p(R, \psi|A, \phi)$ is

$$p(R, \psi|A, \phi) = \sum_{m=1}^{M} \frac{w_m R}{2\pi\sigma^2_{N,m}} \exp\left[-\frac{R^2 + A^2 - 2RA\cos(\psi - \phi)}{2\sigma^2_{N,m}}\right] \tag{6.12}$$

Substituting the likelihood from the equation above and the 1 sided Chi priors from equation 3.21 into the expression for the MMSE estimator given in equation 3.22 and following the same procedure as in appendix A.6 we can show that the resulting estimator is

$$\hat{A} = \frac{\sum_{m=1}^{M} w_m \left(2\sigma^2_{N,m}\right)^{\frac{a-1}{2}} \exp\left[\frac{-R^2}{2\sigma^2_{N,m}}\right] \Gamma\left(\frac{a+1}{2}\right) \zeta_m^{\frac{a+1}{2}} {}_1F_1\left[\frac{a+1}{2}, 1, \frac{R^2\zeta_m}{2\sigma^2_{N,m}}\right]}{\sum_{m=1}^{M} w_m \left(2\sigma^2_{N,m}\right)^{\frac{a-2}{2}} \exp\left[\frac{-R^2}{2\sigma^2_{N,m}}\right] \Gamma\left(\frac{a}{2}\right) \zeta_m^{\frac{a}{2}} \quad {}_1F_1\left[\frac{a}{2}, 1, \frac{R^2\zeta_m}{2\sigma^2_{N,m}}\right]} \tag{6.13}$$

where $\zeta_m = \frac{\theta}{2\sigma^2_{N,m}+\theta}$.

The proposed estimator is compared with the MMSE amplitude estimator with Chi priors and a single Gaussian noise model (MS1C, eq. 3.23). The latter estimator however, allows the noise power to vary with time, to compensate for the time

2-state WGN			
	SegSNR [dB]		
Input	0	10	20
GMM	5.08	11.20	18.93
MS1C	6.14	12.55	20.42
	PESQ		
Input	2.13	2.75	3.41
GMM	2.61	3.17	3.63
MS1C	2.72	3.24	3.70

Train noise			
	SegSNR [dB]		
Input	0	10	20
GMM	3.23	10.36	18.59
MS1C	5.77	12.64	20.73
	PESQ		
Input	2.46	3.04	3.64
GMM	2.65	3.21	3.71
MS1C	2.88	3.35	3.82

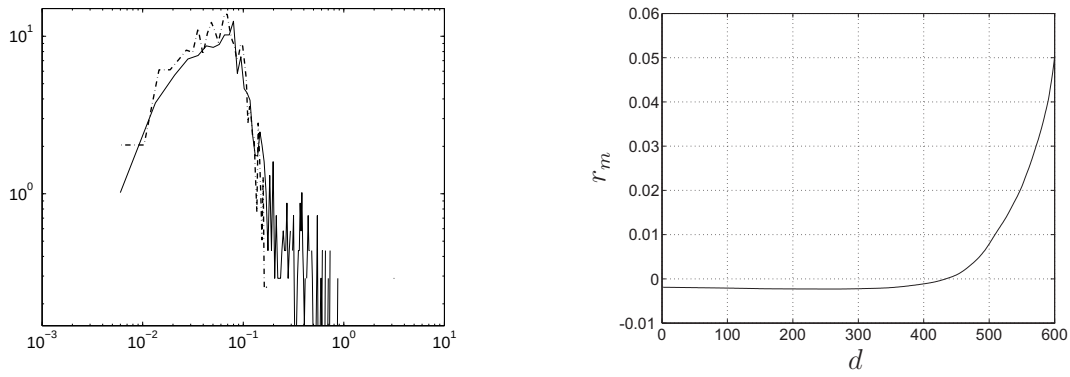
Table 6.1: Comparison of the proposed GMM-based algorithm with the MS1C

variations of the noise signal. The speech signals we are using for the evaluation are 4 sentences from the TIMIT database, two of which are uttered by a male and two by a female. Two different noise signals are used: the first is a stationary white Gaussian noise, whose power increases by 6 dB after the two first utterances. The second is noise recorded in a train, which contains a number of time varying events. The parameters of the GMM model (eq. 6.10) are estimated with the EM algorithm. The noise power estimate that is required for the second algorithm is estimated directly from the noise samples for the first noise signal, while for the second, the power is calculated from the noise signal with a first order recursion of the form

$$\mathcal{P}(k,l) = 0.9\mathcal{P}(k,l-1) + 0.1|N(k,l)|^2 \tag{6.14}$$

The priors' shape parameter a is set to 2 for both algorithms. The results from the above comparison are shown in table 6.1.

Table 6.1 shows that the MS1C algorithm outperforms the algorithm with the GMM noise model. The MS1C algorithm is actually optimal for the two discrete states of the first noise signal. The GMM based algorithm can be considered as optimal for the combination of the two states. The lack of time information about the transition of the states however, poses a clear drawback for the latter algorithm. Even in the case of the train noise, where the transition between the different noise states is less apparent, allowing the noise power to change with time yields better results than using a more accurate but fixed model for the entire duration of the noise segment.



(a) (b)

Figure 6.3: (a) Distributions of the noisy speech (continuous) and noise (dash dot) spectral amplitudes, (b) value of the r_m criterion as a function of the number of samples d of the vector Ω.

6.3 Noise estimation based on matching the moments of the Rayleigh distribution

6.3.1 The Rayleigh Moment Matching noise estimation method

In this section we present an algorithm for the estimation of the noise power, which is also capable of tracking its changes with time. The proposed algorithm, which was published in [5], is based on the similarities between the distribution of consecutive noisy speech STFT amplitude samples within a single frequency bin and the corresponding distribution of the corrupting noise. Figure 6.3(a) shows a typical example of the above distributions. The distributions shown were created from 600 consecutive samples taken from the frequency bin centered at 1 KHz. The clean speech was corrupted with white Gaussian noise at 10 dB SegSNR. An examination of figure 6.3(a) reveals that the leftmost part of the noisy speech distribution resembles the distribution of noise, which is the Rayleigh[2] distribution according to the complex Gaussian model of its STFT coefficients. Additionally, the noisy speech distribution has a much longer tail, which is the result of the high amplitude spectral components that belong to speech. The shape of the noisy speech distribution is the result of the relative sparseness of the speech components within a window of several time frames.

[2]Recall that the Rayleigh distribution is a special case of the 1 sided Chi distribution (eq. 3.21) with $a = 2$.

The above similarities were also exploited by Hirsch and Ehrlicher [48], although that method required the calculation of histograms for every STFT sample of the speech utterance, which resulted in an increased computational load. Our method is based on the same observations as those of Hirsch and Ehrlicher but circumvents the need for the calculation of histograms. The objective we are trying to achieve is the separation of the Rayleigh-like lower part of the noisy speech distribution from its heavy tails. The separation is based on matching the moments of the Rayleigh distribution. The central idea is that if we have a vector of D past noisy spectral amplitude samples and start discarding those with the higher values until the two first moments of the Rayleigh distribution match, then the variance of the remaining samples should give us an estimate of the noise power. The moments are calculated from the data that remain after discarding those with the higher values.

The Rayleigh distribution is given by

$$p(x) = \frac{2x}{\sigma_x^2} \exp\left(-\frac{x^2}{\sigma_x^2}\right), \qquad x \geq 0 \tag{6.15}$$

and its two first moments are $\mathrm{E}[x] = 0.5\sqrt{\pi\sigma_x^2}$ and $\mathrm{E}[x^2] = \sigma_x^2$. Defining a vector of D past spectral amplitude values as $\mathcal{Q} \stackrel{\triangle}{=} \{R(k,l') : l' \in (l-D, l]\}$, which is also *sorted* in ascending order, we can then form the following criterion that can indicate the matching of the two first moments of the Rayleigh distribution

$$r_m(d) = 0.5\sqrt{\pi \mathrm{E}\left[\mathcal{Q}^2(1:d)\right]} - \mathrm{E}\left[\mathcal{Q}(1:d)\right], \qquad d \in [1, D] \tag{6.16}$$

The notation $\mathcal{Q}(1:d)$ indicates the d first samples of the vector $\mathcal{Q}$. If the elements of $\mathcal{Q}$ are drawn from a Rayleigh distribution then the criterion r_m is zero. However, if the vector $\mathcal{Q}$ consists of noisy speech spectral samples the above criterion is typically positive for $d = D$. Decreasing the value of d, until we find a value d_m such that $r_m(d_m) \approx 0$, an estimate for the noise power can then be calculated as $\mathcal{P}(k,l) = \mathrm{E}\left[\mathcal{Q}^2(1:d_m)\right]$. A typical behaviour of r_m as a function of d is shown in figure 6.3(b).

The above procedure has to be repeated for every STFT sample of the speech utterance. However, there is no need for sorting the vector $\mathcal{Q}$ each time. As the analysis progresses by one time frame, it suffices to remove the $R(k, l-D)$ sample

```
For all frequency bins k
        For all time frames l
                Remove R(k, l − D) from Q
                Sort R(k, l) in Q
                If r_m(d_{m|l−1}) > 0
                        Decrease d = d_{m|l−1} until r_m(d) < 0
                else
                        Increase d = d_{m|l−1} until r_m(d) > 0
                P(k, l) = E[Q²(1 : d)],   d_{m|l} = d
```

Table 6.2: Pseudo code for the RMM algorithm

from the vector $\mathcal{Q}$ and sort only the new sample $R(k,l)$. Additionally, if the value of d_m from frame $l-1$ is used as an initial estimate for d in the frame l, the number of evaluations of the r_m criterion can be kept to a minimum. A pseudo code for the proposed algorithm is shown in table 6.2.

6.3.2 Evaluation

We evaluate the performance of the proposed noise estimation algorithm using two tests. In the first, we investigate its ability to track time varying noise, while in the second, the noise estimation algorithm combined with a speech enhancement algorithm is used for the enhancement of speech corrupted by different types of stationary and time varying noise. In both tests the results of the RMM algorithm are compared with those of the Minimum Statistics (MinS) algorithm [71], which is a widely acclaimed algorithm and has often acted as a benchmark for the evaluation of other noise estimation algorithms e.g. [19, 83]. In all the simulations we used $D = 100$ for the RMM algorithm, which was equal to the length of the minima searching window of the MinS algorithm.

For the first test we corrupted four sentences from the TIMIT database with two different types of time varying noise. The first noise was white Gaussian, whose power increased by 6 dB in approximately 3 seconds. The overall input SegSNR was 0 dB. The second noise was recorded in a train and contained some time varying events in the middle of the segment, which were possibly a consequence of the train entering a tunnel. The overall input SegSNR was again 0 dB. The transformation to

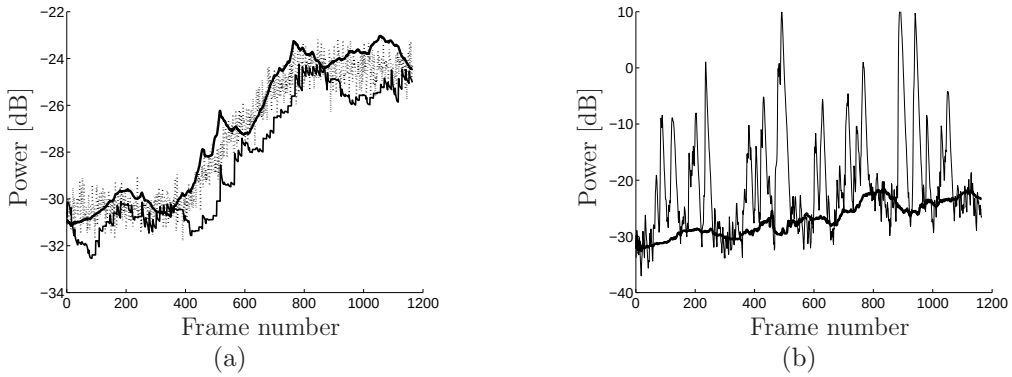


Figure 6.4: Noise tracking results for white Gaussian noise of increasing power. (a) Actual noise power (dashed line), RMM estimate (thick line), MinS estimate (fine line), all averaged across frequency. (b) Smoothed power of noisy speech (fine line) and RMM estimate (thick line) for the frequency bin centered at 1 KHz.

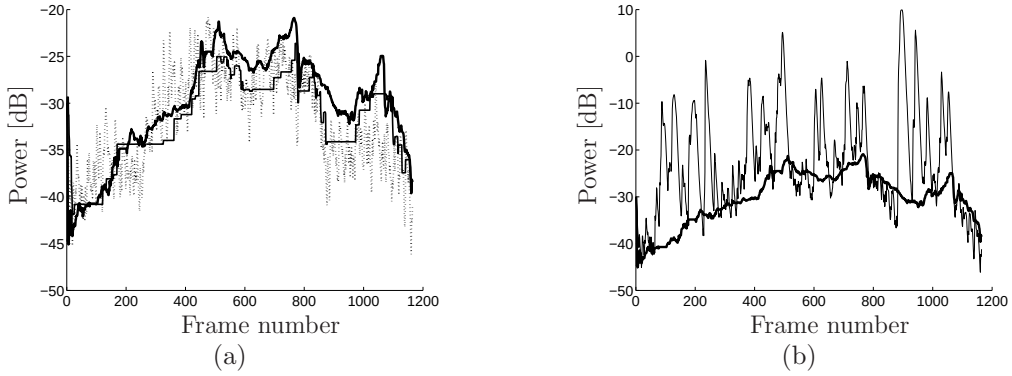


Figure 6.5: Noise tracking results for train noise. (a) Actual noise power (dashed line), RMM estimate (thick line), MinS estimate (fine line), for the frequency bin centered at 1 KHz. (b) Smoothed power of noisy speech (fine line) and RMM estimate (thick line) for the same frequency bin.

the STFT domain was performed with Hamming windows of 256 samples and 75% overlap. Figure 6.4(a) shows the actual power of the white noise (dashed line), the RMM (thick line) and MinS (fine line) estimates of the noise power, all averaged across frequency. Figure 6.4(b) shows the smoothed periodogram of the noisy speech and the RMM noise power estimate for the frequency bin centered at 1 KHz. Figure 6.5(a) shows the smoothed train noise power at the frequency bin centered at 1 KHz and the respective estimates of the RMM (thick line) and MinS (fine line) algorithms. Finally, figure 6.5(b) shows the smoothed noisy speech power for the same frequency bin and the RMM estimate.

Figure 6.4(a) highlights the main advantage of the RMM algorithm, which is the quick response in the event of an increase in the noise power. Observe for example that between frames 400 and 800 the RMM estimates are significantly closer to the actual noise power than the MinS estimates, which are clearly lower. The drawbacks of the RMM algorithm however, involve a slower response in the event of a drop in the noise power (figure 6.5(a) frames 800-900) and a greater tendency to overestimate the noise in the presence of large speech activity (figure 6.5(a), peaks at frames 750 and 1050). The differences in the behaviour of the two algorithms should be attributed to the fact that while the MinS uses only one sample (the minimum) from the window of past spectral values to estimate the noise power, the RMM algorithm essentially employs the d_m minimum values. This makes the RMM more prone to overestimate the noise power in periods of increased speech activity, but also shortens the response time in the event of an increase in the noise power.

For the second evaluation test we corrupted 16 speech sentences from the TIMIT database with stationary white Gaussian noise and with train noise that contained a number of time varying events. The noise estimates provided by the RMM and MinS algorithms were then supplied to the MS1C algorithm with $a = 2$ (a.k.a. Ephraim-Malah MMSE-STSA [31]) in order to evaluate the quality of the resulting speech. The objective evaluation was performed with the SegSNR and the PESQ measures. The results are shown in tables 6.3 and 6.4.

	Noisy	RMM	MinS		Noisy	RMM	MinS
SegSNR	0	7.7	7.7	PESQ	1.62	2.43	2.39
	10	12.7	12.9		2.30	3.04	3.00
	20	16.8	17.2		2.96	3.57	3.56

Table 6.3: White noise results

	Noisy	RMM	MinS		Noisy	RMM	MinS
SegSNR	0	6.6	6.2	PESQ	2.03	2.59	2.53
	10	12.0	12.2		2.67	3.21	3.19
	20	16.8	17.0		3.31	3.70	3.70

Table 6.4: Train noise results

A general trend that can be identified in these tables is that the enhanced speech obtained with the MinS noise estimates scores higher in the SegSNR measure, while

the enhancement algorithm that employed the RMM noise estimates yielded higher PESQ scores. An inspection of enhanced speech spectrograms reveals that the weaker speech spectral components are better preserved with the MinS algorithm. On the other hand, the residual noise exhibits less spurious spectral peaks and a lower overall level when the RMM algorithm is used. This is due to the RMM noise estimates being generally higher than those of the MinS algorithm. Informal listening tests suggest that the better restoration of the weaker speech components offered by the MinS algorithm is less perceptible compared to the more uniform residual noise that is obtained when the RMM algorithm is used, which could also justify the higher PESQ scores.

6.4 Summary

This chapter presented our work on the development of noise estimation algorithms. After summarising the most prominent approaches for noise estimation we presented an alternative modelling approach, which was based on GMM's. The motivation was that GMM's were capable of modelling very accurately the distribution of time varying noise STFT coefficients. The results showed however, that a model based on a single Gaussian distribution might be preferable, as long as the model allows for changes in the variance of the Gaussian distribution with time.

Such an algorithm was presented at the second part of the chapter. The proposed algorithm was based on the similarities between the distribution of the STFT amplitude coefficients of noisy speech, with the distribution of the respective coefficients that belonged to the corrupting noise. An estimate of the noise power was then extracted from the coefficients of a window of past spectral amplitude values that were classified as containing noise only. The classification was performed with a criterion, which was based on matching the two first moments of the Rayleigh distribution. The proposed algorithm exhibited a quick to response in the event of an increase in the noise power, and despite its susceptibility to overestimate the noise power in prolonged periods of speech activity, its overall performance was comparable with that of a state of the art noise estimation method.

Chapter 7

Speech enhancement based on Markov Random Fields

The STFT matrices of speech are known to have a particularly rich structure. Consecutive samples within a frequency bin are highly correlated, as it was demonstrated by Cohen [21]. Additionally, correlations exist between the amplitudes of adjacent frequency bins, which stem not only from the spectral leakage caused by the windows used in the calculation of the STFT, but are also due to the common modulation of the STFT amplitude coefficients in neighbouring frequency bins [3]. Furthermore, the voiced time frames very often have a well defined structure, because of the harmonics of the pitch frequency. The information that is encapsulated in the above attributes of a speech STFT matrix can prove very helpful in the restoration of speech degraded by background noise.

For example, consider the DD method, which is used for the estimation of the a priori SNR, and is popular for its ability to reduce the level of the background noise, and perhaps more importantly to help in the suppression of the musical noise spectral peaks [16]. Equation 4.6 clearly demonstrates the Markovian character of the DD method, and the influence the samples from the previous STFT frame exert on those of the current. The correlation between successive STFT samples has also been exploited by Cohen [21] for the estimation of the a priori SNR, while it is also the main motivation for using Kalman filters for the restoration of the DFT trajectories [102].

In this chapter we present an effort to extend the more traditional unidimensional speech models into both dimensions of the STFT, by exploiting the correlation of speech both in time and in frequency. The framework within which we build our two dimensional time frequency models is provided by the theory of Markov Random Fields (MRF's). The MRF's are spatial stochastic processes, which can be considered as two dimensional extensions of the Markov Chains. Therefore, as the value of a r.v. in a Markov Chain depends on the values of the r.v.'s that precede it, the value of a r.v. in an MRF depends on the values the r.v.'s which are considered as its neighbours, in a two dimensional space.

One of the first studies on MRF's was presented in Besag's seminal paper [10], where the MRF's were rigorously defined and proof was given for some of their fundamental properties. The applications presented in [10] concerned two dimensional agricultural data. The MRF's were brought to the attention of the image and signal processing community with a paper by Geman and Geman [39], where MRF's were employed in the Bayesian restoration of images. Since then, the MRF's have been extensively used in the image processing literature [11, 14, 35, 57, 58, 77, 80].

In speech processing on the other hand, MRF's have found limited applications so far. We are aware of Gravier's work [43], who has employed MRF's in speech recognition, while Andia [2] has tried to tackle the restoration of missing STFT data due to severe contamination from tonal noises. To the best of our knowledge, in the current work it is the first time that Markov Random Fields are used in enhancing speech that has been corrupted with broadband noise.

We begin this chapter by laying down the theoretical background of the Markov Random Fields (§7.1) and introducing the fundamental concepts for their development. A speech enhancement scheme based on Gaussian MRF's is presented in §7.2, which serves as a first example for demonstrating the MRF's ability to incorporate time and frequency dependencies in the estimation model. The Gaussian MRF estimator is unfortunately not well defined for all the values of its input parameters, in a similar fashion as the MP1C and MP1G algorithms of chapter 3. This problem is sidestepped with the introduction of the Chi MRF's in §7.3. The chapter closes with the introduction of an adaptive algorithm that uses the Chi MRF conditional priors and combines a superior restoration of weak speech spectral components with

an effective suppression of the residual noise.

7.1 Theoretical background

In this section we present the basic theory and some fundamental concepts of the Markov Random Fields. Our presentation, which is primarily based on [10] and to some extent on [11, 14, 80], is focused on those aspects of the MRF theory which we will need for developing the proposed speech enhancement schemes in the subsequent sections. A more extensive treatment of the theory of MRF's can be found in the above references.

7.1.1 Markov Random Fields and the Hammersley-Clifford theorem

Suppose that we have a vector of random variables $X = [X_1, ..., X_q]$ and let $x = [x_1, ..., x_q]$ denote a realisation of X. We define the space $\mathcal{S}_i$ of the random variable X_i as

$$\mathcal{S}_i = \{x_i : p(x_i) > 0\}, \quad \text{with} \quad i \in Q = \{1, .., q\}$$

where $p(x_i)$ is the probability density function of X_i. Let also denote the joint probability density function of the X_i random variables as $p(x) = p(x_1, ..., x_q)$. The space $\mathcal{S}$ of the vector of random variables X is given by the Cartesian product of the individual $\mathcal{S}_i$'s

$$\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times, ..., \times \mathcal{S}_q$$

A central concept in the development of Markov Random Fields is that of a *neighbour*. Given two random variables X_i and X_j with $i \neq j$, we say that X_j is a neighbour of X_i if and only if the conditional distribution $p(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_q)$ depends on x_j. The neighbours of the random variable X_i are denoted by $X_{\mathrm{n}(i)}$. We also require that if the realisations $X_1 = x_1, ..., X_q = x_q$ can occur individually, they can also occur simultaneously. More formally, if $p(x_i) > 0 \ \forall i \in Q$ then $p(x_1, ..., x_q) > 0$. The last condition is called the *positivity* condition and is usually satisfied in practice.

Definition. A Markov Random Field is a collection of interacting random variables

with joint probability density function $p(x)$ for which:

(i) The positivity condition holds.

(ii) For each X_i there is a defined set of r.v.'s $X_{\mathrm{n}(i)}$, which are called neighbours and the following statement is true

$$p(x_i|x_{Q-i}) = p(x_i|x_{\mathrm{n}(i)}) \quad \forall i \in Q$$

where, $\{Q - i\}$ is a shorthand notation for the set of indices $j \in Q$ with $j \neq i$.

An intuitively appealing method of constructing an MRF is via the conditional density functions. This method allows the explicit definition of the interactions between a random variable and its neighbours, which is not as straightforward to achieve with a direct construction of a joint density function. The conditional density approach however, is hindered by the disadvantage that not all conditional densities yield a valid joint distribution for the process. This can be illustrated in the following example. Suppose that the vector X consists of two only variables $X = [X_1, X_2]$ and x and z are two different realisations. Bayes' theorem allows us to write the two following expressions assuming that all the conditionals are positive

$$\frac{p(x_1, x_2)}{p(z_1, x_2)} = \frac{p(x_1|x_2)}{p(z_1|x_2)}, \qquad \frac{p(z_1, x_2)}{p(z_1, z_2)} = \frac{p(x_2|z_1)}{p(z_2|z_1)}$$

Multiplication of the above two equations yields

$$\frac{p(x_1, x_2)}{p(z_1, z_2)} = \frac{p(x_1|x_2)}{p(z_1|x_2)} \frac{p(x_2|z_1)}{p(z_2|z_1)} \tag{7.1}$$

An alternative set of expressions could be

$$\frac{p(x_1, x_2)}{p(x_1, z_2)} = \frac{p(x_2|x_1)}{p(z_2|x_1)}, \qquad \frac{p(x_1, z_2)}{p(z_1, z_2)} = \frac{p(x_1|z_2)}{p(z_1|z_2)}$$

and consequently

$$\frac{p(x_1, x_2)}{p(z_1, z_2)} = \frac{p(x_2|x_1)}{p(z_2|x_1)} \frac{p(x_1|z_2)}{p(z_1|z_2)} \tag{7.2}$$

There is no obvious reason why the right hand sides of eqs. 7.1 and 7.2 should be equal, which implies that there must be some 'hidden' constraints in the form of the conditional densities that result in valid joint density functions.

The above constraints are explored by the *Hammersley-Clifford* theorem [10, 17, 44], which poses the question: *given the neighbours of each r.v. and the positivity condition, what is the most general form the joint density $p(x)$ can take in order to define a valid probability structure to the system.* Assuming that x_i^0 is a realisation of X_i and defining $x^0 \equiv [x_1, ..., x_{i-1}, x_i^0, x_{i+1}, ..., x_q]$ we have

$$\frac{p(x)}{p(x^0)} = \frac{p(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_q)}{p(x_i^0|x_1, ..., x_{i-1}, x_{i+1}, ..., x_q)} \tag{7.3}$$

The above equation implies that knowledge of the most general form $p(x)$ can take, provides the most general form for the conditional densities as well.

Theorem (Hammersley-Clifford) Let $\{p(x) > 0 : x \in \mathcal{S}\}$, denote a probability density function satisfying the positivity condition. Then p(x) is a Markov Random Field if and only if

$$p(x) \propto \prod_{\mathcal{C}} \Psi_{\mathcal{C}}(x_{\mathcal{C}}) \tag{7.4}$$

where the functions $\Psi_{\mathcal{C}}(x_{\mathcal{C}})$ are chosen arbitrarily, subject to $0 < \Psi_{\mathcal{C}}(x_{\mathcal{C}}) < \infty$ for all $x \in \mathcal{S}$. The sets of indices $\mathcal{C} \subseteq Q$ define sets of random variables $x_{\mathcal{C}}$, which in the MRF literature are termed *cliques*. A clique is a set in which every random variable is a neighbour of every other random variable in that set. Cliques can also be sets that consist of a single random variable (singleton).

The above theorem was originally derived for discrete random variables, where $p(x)$ denoted a probability *mass* function. Its extension to continuous random variables was however straightforward, subject to the integrability of $p(x)$, which then denoted a probability *density* function.

Although a number of different neighbourhood schemes exists [10] we will only be concerned with first order schemes, as the one depicted in figure 7.1. In this scheme, the random variables are arranged on a rectangular lattice and each one of them depends on the values of its four nearest neighbours. As the lattice is not infinite, the r.v's at its edges will only have three neighbours, while the r.v.'s at the corners will only have two. The cliques in this spatial scheme consist only of singletons and pairs of neighbours. Therefore, only pairwise interactions are allowed between the random variables.

.	.	.	.	.
.	.	x	.	.
.	x	o	x	.
.	.	x	.	.
.	.	.	.	.

Figure 7.1: First order MRF neighbourhood. The cells that contain the 'x' are the neighbours of 'o'. The distribution of 'o' is independent of all the other cells if the values of the 'x's are known.

7.1.2 Gaussian Markov Random Fields

A type of continuous MRF that is very common and serves as a good introductory application of the theoretical developments of the previous section is the Gaussian MRF. Its conditional density function is

$$p(x_i|x_{\mathrm{n}(i)}) \propto \exp\left[-\frac{1}{2\sigma_i^2}\left(x_i - \sum_{j\in \mathrm{n}(i)} b_{ij}x_j\right)^2\right] \tag{7.5}$$

where σ_i^2 is the variance of x_i and b_{ij} is a weight that determines the influence of x_j on x_i. The joint density function can be derived via the factorisation

$$\frac{p(x)}{p(z)} = \prod_{i\in Q} \frac{p(x_i|x_1, ..., x_{i-1}, z_{i+1}, ..., z_q)}{p(z_i|x_1, ..., x_{i-1}, z_{i+1}, ..., z_q)} \tag{7.6}$$

Substituting the expression for the conditional density from eq. 7.5 in the above factorisation, and assuming the symmetry condition $b_{ij}/\sigma_i^2 = b_{ji}/\sigma_j^2$ the derived expression for the joint density is [1]

$$p(x) \propto \exp\left[-\left(\sum_{i\in Q}\frac{x_i^2}{2\sigma_i^2} - \sum_{\{i,j\}\in C}\frac{b_{ij}}{\sigma_i^2}x_i x_j\right)\right] \tag{7.7}$$

where C denotes the unordered set of pairs of indices, such that $\{i, j\} \in C$ if and only if x_i and x_j are neighbours. Note also that $x_{\mathcal{C}} \equiv [x_Q, x_C]$, or in other words the cliques in a first order neighbourhood consist of the r.v.'s that form the MRF and the pairs of r.v.'s which are mutually neighbours.

[1] see also appendix D for a similar derivation of the Chi MRF joint density, which is a generalisation of the Gaussian MRF.

The density in eq. 7.7 can be written as

$$p(x) \propto \exp\left[-xGx^{\mathrm{T}}\right] \tag{7.8}$$

where G is a matrix with elements $G_{ii} = 1/2\sigma_i^2$ and $G_{ij} = -b_{ij}/2\sigma_i^2$. The above density function corresponds to the multivariate Gaussian. We see therefore, that the conditionally Gaussian MRF leads to a multivariate Gaussian density. A condition for the above density to be valid is that the matrix G is positive definite. Bouman and Sauer [14] state that a sufficient condition for the positive definiteness of G is that all of its elements are positive ($G_{ij} > 0 \forall i, j \in Q$) and that $G_{ii} > \sum_{j \in \mathrm{n}(i)} G_{ij}$, $\forall i \in Q$. A proof of the last statement is given in appendix D.

7.1.3 Estimation with MRF priors

Suppose that we observe a set of random variables $Y = [Y_1, ..., Y_q]$, which are modelled as a random function of the random variables X that constitute an MRF. An example of such a random function could be the addition of a Gaussian noise vector to X. We additionally suppose that the random variables Y are independent when the values of X are given. For the joint density function of Y we therefore have

$$p(y) = \prod_{i \in Q} p(y_i|x_i) \tag{7.9}$$

A typical estimation problem under the above scenario is to find an optimal, in some sense, estimate of X when only Y is observed, given that the joint density of X belongs to the class of Markov Random Fields.

An estimator that has been widely used in the literature is the MAP, which according to Bayes' rule it can be written as (see §2.3.2)

$$\hat{x} = \arg\max_x p(y|x)p(x) \tag{7.10}$$

The above optimisation problem can be enormously difficult to solve due to the typically large number of the random variables involved in real problems. In an image processing scenario for example, even a small picture (256×256) contains 2^{16}

pixels. A relatively efficient, although still demanding computationally optimisation method, was proposed by Geman and Geman [39] involving simulated annealing and the Gibbs sampler. Apart from the heavy computational load, an additional disadvantage of this type of global optimisation is that it can induce positive correlations between random variables that are arbitrary far from each other [11], while it is generally desirable to have models whose dependencies are only local.

An alternative local instead of global optimisation method was proposed by Besag [11], which was termed Iterated Conditional Modes (ICM). In this estimation scheme the proposed estimate $\hat{x}_i$ is the one with the maximum probability given the observation y_i and its neighbours $x_{\mathrm{n}(i)}$. That is,

$$\hat{x}_i = \arg\max_{x_i} p(y_i|x_i)p(x_i|x_{\mathrm{n}(i)}) \tag{7.11}$$

The ICM method circumvents the problems posed by the computational load of the global MAP estimate and the large scale dependencies. However, Besag [11] states that it has no proper mathematical basis, mainly due to the reason that it does not always converge to the global MAP solution, which is the theoretically sound solution according to the MRF model specifications. Nevertheless, the computational efficiency and the mitigation of large scale dependencies offered by the ICM, question the need for a strict adherence to the MRF theory, taking also into account the fact that the ICM method does not require the conditional distributions to define a legitimate MRF joint density, as predicted by the Hammersley-Clifford theorem. In Besag's words 'it is only a partial answer to the above question to suggest that adherence to genuine MRF's removes some arbitrariness and aids interpretation' [11].

7.2 Speech enhancement based on Gaussian MRF priors

In this section we propose a speech enhancement scheme that uses a Gaussian Markov Random Field as a model of the speech spectral amplitude samples. In particular, a MAP estimator of the speech spectral amplitude is derived using a Gaussian MRF prior. The estimation method used is the ICM, because of its low computational load and because we wish to incorporate in our model only local,

and not large scale, interactions between the spectral amplitude samples. We begin by deriving the proposed estimator, then define the neighbourhood and finally, discuss the implementation of the estimator. The evaluation of the proposed scheme is given in §7.2.4. A version of the algorithm proposed in this section has also been published in [6].

7.2.1 Derivation of GMRF the estimator

The derivation of this estimator is very similar to that of the MP1C (§3.3.1.2 and appendix A.7). The difference between the two estimators is that the MRF prior for the spectral amplitude sample A_i is conditioned on its neighbours (i.e. $p(A_i|A_{\mathrm{n}(i)})$), while the Chi prior (eq. 3.21), used in the MP1C estimator, was a function of the sample A_i alone (i.e. $p(A_i)$). The proposed estimator can be found by maximising the expression

$$\hat{A}_i = \arg\max_{A_i} \ln\left[p(R_i|A_i)p(A_i|A_{\mathrm{n}(i)})\right] \tag{7.12}$$

where the Gaussian MRF prior is given by

$$p(A_i|A_{\mathrm{n}(i)}) \propto \exp\left[-\frac{1}{2\sigma_i^2}\left(A_i - \sum_{j\in\mathrm{n}(i)} b_{ij}A_j\right)^2\right] \tag{7.13}$$

In the above expression σ_i^2 represents the variance of the sample A_i and b_{ij} is the weight between A_i and A_j. To obtain the likelihood $p(R_i|A_i)$ we use the approximate expression, derived in appendix A.1 eq. A.14

$$p(R_i|A_i) \propto A_i^{-1/2} \exp\left[-\frac{R_i^2 + A_i^2}{2\sigma_{N,i}^2}\right] \exp\left[\frac{R_i A_i}{\sigma_{N,i}^2}\right] \tag{7.14}$$

which allows the derivation of the estimator in closed form, as was the case with the MP1C and MP1G estimators. In the above equation $2\sigma_{N,i}^2 \equiv \mathrm{E}[|\mathbf{N}_i|^2]$ where $\mathbf{N}_i$ is the i^{th} sample of the (complex) noise STFT (eq. 3.2).

The expression for the estimator can then be written as

$$\hat{A}_i = \arg\max_{A_i}\left[-\frac{\ln(A_i)}{2} - \frac{(A_i - R_i)^2}{2\sigma_{N,i}^2} - \frac{\left(A_i - \sum_{j\in\mathrm{n}(i)} b_{ij}A_j\right)^2}{2\sigma_i^2}\right] \tag{7.15}$$

Differentiating the above expression w.r.t. A_i and setting the result equal to zero, the MAP estimate for A_i can be expressed as

$$\hat{A}_i = \zeta_1 + \sqrt{\zeta_1^2 - \zeta_2} \tag{7.16}$$

where

$$\zeta_1 = \frac{R_i \sigma_i^2 + \sigma_{N,i}^2 \sum_{j \in \mathrm{n}(i)} b_{ij} A_j}{2\left(\sigma_i^2 + \sigma_{N,i}^2\right)}, \qquad \zeta_2 = \frac{\sigma_{N,i}^2 \sigma_i^2}{2\left(\sigma_{N,i}^2 + \sigma_i^2\right)}$$

The above estimator with $b_{ij} = 0$ yields the MP1C estimator with shape parameter $a = 1$ (half Gaussian priors).

Similar to the MP1C estimator with $a = 1$, the GMRF estimator (eq. 7.16) is not well defined for all the values of its input parameters, as it can be seen from the discriminant in eq. 7.16, which can take negative values. This is a consequence of the approximation of the Bessel function (eq. A.13), which introduces a singularity in the likelihood (eq. 7.14) for $A_i = 0$. As was the case with the MP1C algorithm, the estimate of eq. 7.16 is used only when the discriminant is non negative, while for negative values the noisy sample R_i is suppressed by 50 dB.

7.2.2 Definition of the neighbourhood

The selection of the neighbours of the sample A_i determines its interaction with the rest of the spectral amplitude samples. Our motivation for employing the MRF priors was the incorporation of the time and frequency dependencies of speech in the statistical model. Initially, we experimented with a simple four sample neighbourhood, where the neighbours were the immediately adjacent samples of A_i in time and frequency. Adopting the notation $A(k, l)$ to represent the spectral amplitude, where k and l denote the frequency and time indices respectively, such a neighbourhood is defined as

$$A_{\mathrm{n}(k,l)} = \{A(k-1, l), A(k+1, l), A(k, l-1), A(k, l+1)\} \tag{7.17}$$

In the course of this work, it became apparent that a 'harmonic' neighbourhood, similar to the one proposed by Andia [2], provides better results during voiced time frames. In the harmonic neighbourhood the frequency neighbours for the voiced

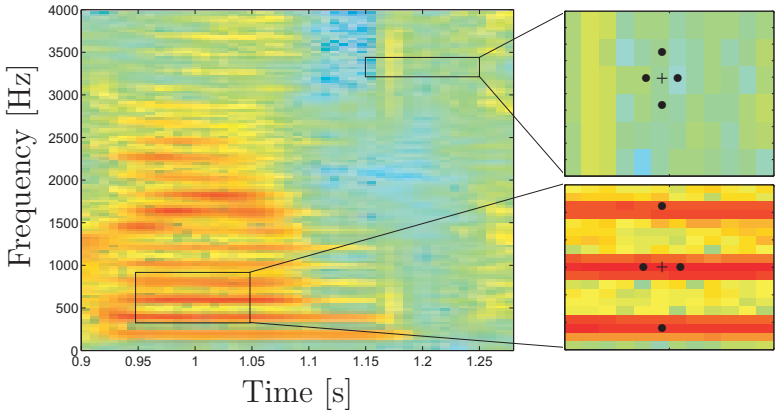


Figure 7.2: Illustration of the proposed harmonic neighbourhood. Upper right figure shows the neighbours of a sample that belongs to an unvoiced frame and lower right figure shows the neighbours used for the samples of the voiced frames.

frames are k_{f0} frequency bins apart, where k_{f0} is the frequency bin number that corresponds to the pitch frequency of the current frame. (Assuming that the DC frequency bin is numbered 0.) The definition of the harmonic neighbourhood is given in eq. 7.18 and is illustrated in figure 7.2.

$$A_{\mathrm{n}(k,l)} = \begin{cases} \{A(k-1,l), A(k+1,l), A(k,l-1), A(k,l+1)\} & \text{if } l \text{ unvoiced} \\ \{A(k-k_{f0},l), A(k+k_{f0},l), A(k,l-1), A(k,l+1)\} & \text{if } l \text{ voiced} \end{cases} \tag{7.18}$$

As a shorthand notation for the neighbours of $A_i \equiv A(k,l)$ we introduce the notation $A_{\mathrm{n}(i)} = \{A_{\mathrm{S}}, A_{\mathrm{N}}, A_{\mathrm{W}}, A_{\mathrm{E}}\}$, which denote the south, north, west and east neighbours respectively. If the frame l is unvoiced we denote $A_{\mathrm{S}} \equiv A(k-1,l)$ and $A_{\mathrm{N}} \equiv A(k+1,l)$, while if frame l is voiced $A_{\mathrm{S}} \equiv A(k-k_{f0},l)$ and $A_{\mathrm{N}} \equiv A(k+k_{f0},l)$. In both cases it holds that $A_{\mathrm{W}} \equiv A(k,l-1)$ and $A_{\mathrm{E}} \equiv A(k,l+1)$[2].

The above type of neighbourhood implies that a pitch estimate for each of the voiced frames is needed. The estimates are obtained with the pitch estimator of the 2400 bps Federal Standard Speech Coder [93]. This pitch estimation algorithm is based on autocorrelation and the application of error correcting procedures for common

[2]The DC and the Nyquist frequency bins are calculated with the weights of the frequency neighbours A_{S}, A_{N} set to zero. The voiced frame samples, which are below the pitch frequency are also calculated with a local neighbourhood. That is because they typically have low speech energy and the local neighbourhood avoids contamination from frequency bins above the pitch frequency, which typically have higher energy. Finally, the samples of the voiced frames that are less than a pitch frequency apart from the Nyquist frequency bin have the north neighbour's weight set to zero.

errors such as pitch doubling.

Regarding the performance of the pitch estimator, we should mention that the proposed speech enhancement algorithm does not require a great accuracy for the pitch estimates. The reason is that only the frequency bin number that corresponds to the pitch frequency is required and not the actual pitch frequency. In our experiments we have used analysis windows of 256 samples for calculating the STFT coefficients, while the sampling frequency was 8 KHz. This implies that each frequency bin corresponds to a bandwidth of 31.25 Hz, which makes the speech enhancement algorithm robust to relatively small inaccuracies of the pitch estimator.

7.2.3 Implementation

In order to obtain an estimate for A_i according to eq. 7.16 there are a number of quantities related to the i^{th} STFT point that must be known (i.e. R_i, $\sigma_{N,i}^2$, σ_i^2), as well as quantities that correspond to neighbouring STFT points, such as A_j and b_{ij}. In this section we discuss the estimation of the above quantities, not all of which are readily available during the estimation of A_i. Additionally, the definition of a valid MRF scheme requires that $b_{ij}/\sigma_i^2 = b_{ji}/\sigma_j^2$ (see §7.1.2). We also explain how the above requirement is fulfilled within our estimation scheme.

The variance of the noise spectral amplitude coefficients $\sigma_{N,i}^2$ can be obtained from a noise estimation algorithm and the value of σ_i^2 can be estimated from the a priori SNR ξ_i. The latter quantity is estimated with the DD method, as shown in chapter 4, while the relationship between σ_i^2 and ξ_i is $\sigma_i^2 = 2\xi_i\,\sigma_{N,i}^2$.

The estimation of the speech spectral amplitudes A_i proceeds first from smaller to larger frequency indices k and subsequently form smaller to larger time frame indices l. According to this estimation 'schedule', during the estimation of A_i estimates exist for A_{S} and A_{W}. The same is not true for A_{E} and A_{N}, although their values are required according to eq. 7.16. For this reason, temporary estimates of A_{E} and A_{N} are calculated, which are used only for the estimation of A_i and are then discarded. The estimates are calculated with eq. 7.16 setting the neighbour weights b_{ij} to zero.

Finally, in order to generate a valid MRF scheme, the symmetry condition $b_{ij}/\sigma_i^2 =$

```
For all time frames l
    For all frequency bins k
        Obtain σ²_{N,i} from the noise estimation algorithm
        Estimate A_N, σ²_N with eq. 7.19 and b_ij = 0
        Estimate A_E, σ²_E with eq. 7.19 and b_ij = 0
        Estimate σ_i = 2ξ_i σ²_{N,i} with the DD method
        Estimate Σ_{j∈n(i)} (b_ij/σ²_i) A_j according to eq. 7.20
        Estimate Â_i according to eq. 7.19
```

Table 7.1: Pseudo code for the GMRF algorithm

b_{ji}/σ_j^2 must be satisfied. A method of achieving this is to first write eq. 7.16 as

$$\hat{A}_i = \zeta_1 + \sqrt{\zeta_1^2 - \zeta_2} \tag{7.19}$$

where

$$\zeta_1 = \frac{R_i + \sigma_{N,i}^2 \sum_{j\in \mathrm{n}(i)} \frac{b_{ij}}{\sigma_i^2} A_j}{2\left(1 + \sigma_{N,i}^2/\sigma_i^2\right)}, \qquad \zeta_2 = \frac{\sigma_{N,i}^2 \sigma_i^2}{2\left(\sigma_{N,i}^2 + \sigma_i^2\right)}$$

Then by setting the summation term of ζ_1 equal to

$$\sum_{j\in \mathrm{n}(i)} \frac{b_{ij}}{\sigma_i^2} A_j = \frac{b_{i\mathrm{S}}}{\sigma_i^2} A_{\mathrm{S}} + \frac{b_{i\mathrm{N}}}{\sigma_{\mathrm{N}}^2} A_{\mathrm{N}} + \frac{b_{i\mathrm{W}}}{\sigma_i^2} A_{\mathrm{W}} + \frac{b_{i\mathrm{E}}}{\sigma_{\mathrm{E}}^2} A_{\mathrm{E}} \tag{7.20}$$

and ensuring that $b_{i\mathrm{S}} = b_{i\mathrm{N}}$ and $b_{i\mathrm{W}} = b_{i\mathrm{E}}$ the symmetry condition is fulfilled. In the above equations σ_{N}^2 and σ_{E}^2 are the second moments of A_{N} and A_{E} respectively, which had already been calculated during the temporary estimation of A_{N} and A_{E}. A pseudo code for the GMRF algorithm is shown in table 7.1

7.2.4 Results

For the evaluation of the proposed GMRF algorithm we use the simulation setup described in §5.1. That is, 48 sentences from the TIMIT database are corrupted with additive white Gaussian and car noise at three different input Segmental SNR's and the noisy sentences are enhanced with the proposed algorithm. The noise power is estimated directly from the noise samples, in order to eliminate the effect of a

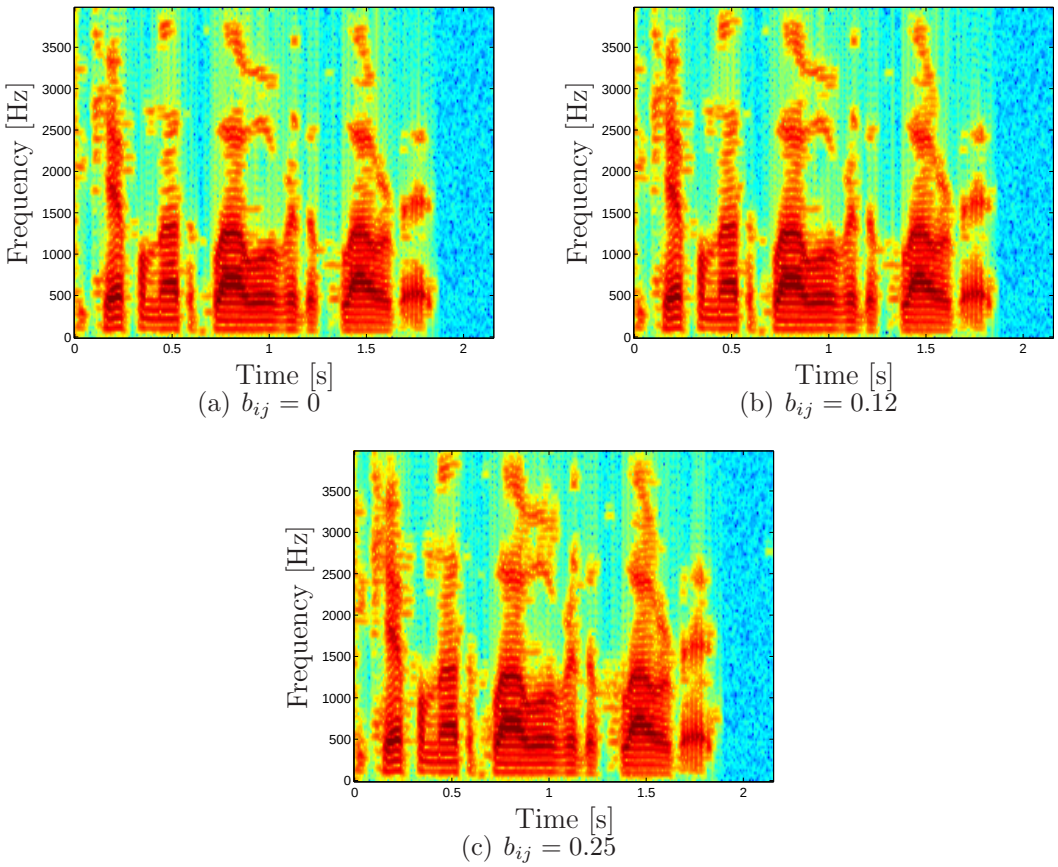


Figure 7.3: Speech enhanced with the GMRF algorithm and different values of b_{ij}.

noise estimation algorithm. The objective speech quality measures used are the SegSNR and the PESQ (§5.2), while the smoothing parameter α of the DD method is set to 0.99. Along with the objective speech quality measures, we also show some spectrograms that illustrate the effect of the neighbour coupling parameter on the enhanced speech. All the spectrograms shown in this chapter correspond to the same phrase used in §5.2 ('*Be careful not to plough over the flower beds*').

Figure 7.3 shows a spectrogram of speech enhanced with the GMRF algorithm and three different neighbour weights b_{ij}. The spectrograms show that as the values of b_{ij} increase more speech spectral components are preserved. This is the result of the time and frequency coupling that the b_{ij} weights impose.

Tables 7.2, 7.3 show the results in the objective measures obtained with the speech database described in §5.1. Although,the SegSNR scores increase with the value of b_{ij}, the PESQ scores remain roughly the same for $b_{ij} = 0$ and 0.12, while they drop for $b_{ij} = 0.25$. The PESQ scores reveal the fact that the excessive coupling

between the neighbours generates estimation artifacts that deteriorate the quality of the enhanced speech. Therefore, although from figure 7.3 the speech harmonics seem to be better preserved when b_{ij} equals 0.25 rather than 0.12, the higher spectral estimates of the former case, particularly in the segments between the words (e.g. between 1.3-1.4 secs), distort the speech rather than enhancing its quality.

	Noisy	$b_{ij} = 0$	$b_{ij} = 0.12$	$b_{ij} = 0.25$
SegSNR	0	7.17	7.51	7.47
	10	12.72	13.25	13.39
	20	20.02	20.72	21.02
PESQ	2.11	2.63	2.61	2.55
	2.80	3.17	3.17	3.14
	3.46	3.80	3.82	3.81

Table 7.2: Objective measure results for white noise

	Noisy	$b_{ij} = 0$	$b_{ij} = 0.12$	$b_{ij} = 0.25$
SegSNR	0	10.71	10.90	10.83
	10	16.59	16.82	16.83
	20	23.76	24.00	24.10
PESQ	2.89	3.33	3.33	3.26
	3.49	3.83	3.83	3.79
	4.07	4.19	4.19	4.18

Table 7.3: Objective measure results for car noise

The most serious drawback of the GMRF algorithm however, stems from the fact that the estimator is not defined when the discriminant in eq. 7.16 is negative. This characteristic, which is also present in the MP1C and MP1G estimators of chapter 3 for small values of a, generates large differences between the estimates obtained for input samples that have a marginally positive or negative discriminant. The result is the appearance of some isolated spectral peaks, which are perceived as musical tones and are unfortunately amplified as the coupling between the neighbours increases. An attempt to rectify this problem, and create a well defined estimator for all the values of its input parameters, will be made in the next section with the introduction of the Chi Markov Random Fields.

7.3 Speech enhancement based on Chi MRF priors

The MP1C estimator of chapter 3 was not well defined for all the values of its input parameters, when the shape parameter a was less than 1.5. This resulted in enhanced speech that suffered from musical noise, due to the fact that the output of the estimator was not continuous with respect to its input arguments. This problem was however rectified when the value of a increased beyond 1.5. We have seen in §7.2.1 that the MP1C algorithm with $a = 1$ is a special case of the GMRF algorithm, obtained by setting $b_{ij} = 0$. Additionally, the GMRF algorithm was ill defined in the same sense as the MP1C with $a = 1$, also suffering from musical residual noise. In this section we make an attempt to rectify these shortcomings of the GMRF algorithm by introducing the Chi MRF priors, whose parameter a can be tuned so that the resulting estimator is well defined for all the values of its input parameters, so the residual noise of the resulting speech has a uniform character.

7.3.1 Chi Markov Random Fields - the CMRF Estimator

The Chi MRF is an extension of the Gaussian MRF in an analogous fashion with the Chi density function being a generalisation of the Gaussian. We define the conditional density function of the Chi MRF as

$$p(A_i|A_{\mathrm{n}(i)}) \propto A_i^{a-1} \exp\left[-\frac{1}{\theta_i}\left(A_i - \sum_{j\in \mathrm{n}(i)} b_{ij}A_j\right)^2\right] \tag{7.21}$$

Under the assumption that $b_{ij}/\theta_i = b_{ji}/\theta_j$, the joint density function for the Chi MRF is (appendix D)

$$p(A) \propto \prod_{i\in Q}\left(A_i^{a-1}\right) \exp\left[-\sum_{i\in Q}\frac{A_i^2}{\theta_i} + \sum_{\{i,j\}\in C}\frac{2b_{ij}}{\theta_i}A_iA_j\right] \tag{7.22}$$

A sufficient condition for the above expression to constitute a valid probability density function (i.e. $|\int_A p(A)dA| < \infty$) is that $b_{ij} > 0,\ \forall i,j \in Q$ and $\sum_{j\in \mathrm{n}(i)} b_{ij} < 1,\ \forall i \in Q$. The above sufficiency condition is proved in appendix D, demonstrating that the Chi MRF's are valid MRF schemes.

The procedure for deriving the estimator based on the Chi MRF priors is identical to the procedure followed for the GMRF. Substituting the expression for the likelihood (eq. 7.14) and the Chi MRF prior (eq. 7.21) in eq. 7.12, yields the following expression for the estimator

$$\hat{A}_i = \arg\max_{A_i} \left[-\frac{\ln(A_i)}{2} - \frac{(A_i - R_i)^2}{2\sigma_{N,i}^2} + \ln(A_i^{(a-1)}) - \frac{\left(A_i - \sum_{j\in \mathrm{n}(i)} b_{ij} A_j\right)^2}{\theta_i} \right] \tag{7.23}$$

The maximum of the above expression can be found by setting its first derivative to zero. The resulting estimator can be expressed as

$$\hat{A}_i = \zeta_1 + \sqrt{\zeta_1^2 - \zeta_2} \tag{7.24}$$

where

$$\zeta_1 = \frac{R_i\theta_i + 2\sigma_{N,i}^2 \sum_{j\in \mathrm{n}(i)} b_{ij} A_j}{2\left(\theta_i + 2\sigma_{N,i}^2\right)}, \qquad \zeta_2 = (1.5 - a)\frac{\sigma_{N,i}^2 \theta_i}{\theta_i + 2\sigma_{N,i}^2}$$

The implementation of the above estimator is identical to that of the GMRF estimator. The same harmonic neighbourhood is also employed. The parameter σ_i^2 of the GMRF estimator has been replaced by θ_i in the CMRF and its value is calculated via the a priori SNR from the relation (see table 4.3)

$$\theta_i = \frac{4\sigma_{N,i}^2 \xi_i}{a}$$

7.3.2 Results

The simulation setup used for the evaluation of the CMRF algorithm is identical to that used for the evaluation of the GMRF algorithm, which was described in §7.2.4. Figure 7.4 shows spectrograms of a speech utterance enhanced with the CMRF algorithm with $a = 1$ and $b_{ij} = 0$, 0.06 and 0.12. The spectrograms reveal that as the coupling imposed by b_{ij} increases more speech spectral components are recovered. Nevertheless, despite the fact that the residual noise has a uniform character, its level increases with increasing values of b_{ij}. Informal listening tests also

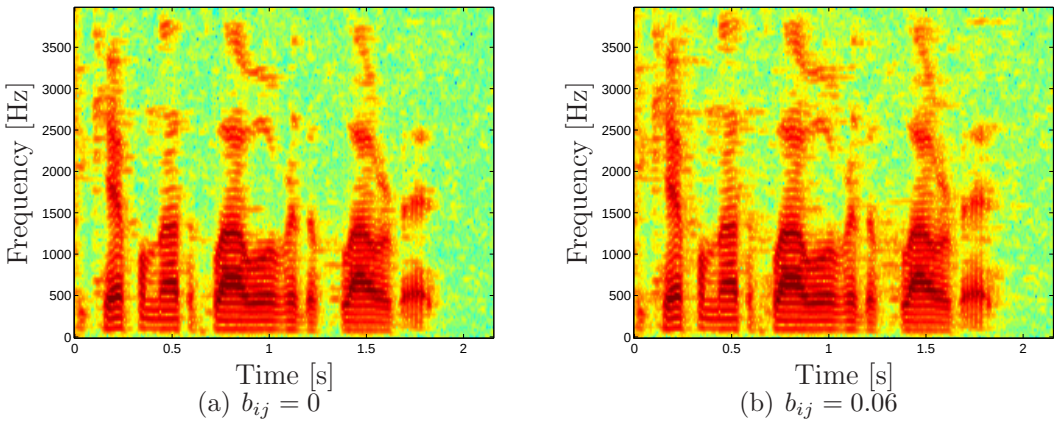


Figure 7.4: Speech enhanced with the CMRF algorithm and different values of b_{ij}.

indicate that the merits of the better preservation of the speech spectral components are outweighed by the increased level of the residual noise for all values of b_{ij}. This is also indicated in the results of the objective measures which drop monotonically with the increase of b_{ij} in almost all cases, the only exception being the SegSNR at the highest input SNR. Tables 7.4, 7.5 present these results for different values of b_{ij}. The results in the above tables were obtained with the speech database described in §7.2.4.

An interpretation as to why the coupling between the neighbours increases the residual noise level can be provided by inspection of the formula of the CMRF estimator (eq. 7.24). Setting $b_{ij} = 0$ yields the MP1C estimator, while $b_{ij} > 0$ implies that

$$\hat{A}_{CMRF} \geq \hat{A}_{MP1C}$$

where $\hat{A}_{CMRF}$ and $\hat{A}_{MP1C}$ are the CMRF and MP1C estimates. As the influence of the neighbours does not depend on whether they contain speech plus noise or

noise only information, since the b_{ij}'s are fixed, increasing the coupling will increase the estimates of the noise only regions of the spectrogram as well as those that contain speech. This effect was not as prominent in the GMRF algorithm, because of the hard thresholding, arising when negative values of the discriminant were encountered, which occurred mainly in the noise dominated portions of the utterance. However, as the CMRF estimator is well defined for all the values of its input arguments the above effect is more pronounced. In the following section we present an adaptive method for the estimation of the neighbour weights that avoids the above problem and enhances the spectral components that contain mostly speech.

	Noisy	$b_{ij} = 0$	$b_{ij} = 0.12$	$b_{ij} = 0.25$
SegSNR	0	6.80	4.92	2.94
	10	12.59	12.38	11.57
	20	20.06	20.74	20.53
PESQ	2.11	2.79	2.57	2.33
	2.80	3.28	3.12	2.97
	3.46	3.77	3.63	3.55

Table 7.4: Objective measure results for white noise

	Noisy	$b_{ij} = 0$	$b_{ij} = 0.12$	$b_{ij} = 0.25$
SegSNR	0	10.20	9.70	7.68
	10	16.53	16.46	15.52
	20	23.86	23.99	23.60
PESQ	2.89	3.36	3.31	3.13
	3.49	3.79	3.76	3.65
	4.07	4.17	4.16	4.12

Table 7.5: Objective measure results for car noise

7.3.3 Adaptive selection of the neighbour weights

In this section we propose the use of adaptive weights between the neighbours, in order to improve the restoration of the speech spectral components, without increasing the level of the residual noise. In order to attain this goal we make the b_{ij} weights a function of the local SNR, so that the influence of a neighbour increases when its SNR is high and vice versa. Additionally, we wish to decouple the parameter θ_i from the a priori SNR and the DD method, in order to avoid having the MP1C estimates as a lower bound of the new algorithm's estimates. In other

words, in the new algorithm, which is referred to as Adaptive CMRF (ACMRF), we wish to remove the constraint $\hat{A}_{ACMRF} \geq \hat{A}_{MP1C}$.

The proposed estimates for the parameters θ_i and b_{ij} are

$$\frac{\theta_i}{2\sigma^2_{N,i}} = \frac{w_{ii}\xi_i^l}{\sum_{m\in \mathrm{n}(i)} w_{im}\xi_m^l + a/2} \tag{7.25}$$

and

$$b_{ij} = \frac{w_{ij}\sqrt{\rho_{ij}\,\xi_j^l}}{\sum_{m\in \mathrm{n}(i)} w_{im}\xi_m^l + a/2} \tag{7.26}$$

The term ξ_i^l is a local a priori SNR at sample i, defined as $\xi_i^l \equiv \frac{\mathrm{E}[A_i^2]}{2\sigma^2_{N,i}}$. The term ρ_{ij} on the other hand is a 'cross' a posteriori SNR between samples i and j defined as $\rho_{ij} \equiv \frac{R_i^2}{2\sigma^2_{N,j}}$. The constants w_{ij} are weights that control the amount of interaction between the neighbours. The expression in eq. 7.25 is essentially a ratio between the SNR of the sample that is being estimated and the SNR of its neighbours. Under this estimation scheme the term $\theta_i/2\sigma^2_{N,i}$ is large when the local SNR at the sample i is higher than that of its neighbours and vice versa. The expression for b_{ij} was chosen with a similar concept in mind, but as the development that leads to the exact form of eq. 7.26 is more involved, it will be presented later in §7.3.5.

We now consider the estimation of the parameters that are involved in eqs. 7.25 and 7.26. We assume the same schedule for the estimation of the STFT samples as in §7.2.3, i.e. first from smaller to larger frequency indices k and subsequently form smaller to larger time frame indices l. According to this, during the estimation of A_i there are available estimates for A_{S} and A_{W}, but not for the A_{N} and A_{E}. For the two latter quantities we need to calculate temporary estimates. The temporary estimate for A_{N} is found as

$$\hat{A}_{\mathrm{N}} = \left(R^2_{\mathrm{N}} - 2\sigma^2_{N,\mathrm{N}}\right)^{0.5} \tag{7.27}$$

and an equivalent formula is used for the estimation of A_{E}. For the a local priori SNR's of the neighbours we propose the estimates

$$\hat{\xi}_j^l = \frac{\hat{A}_j^2}{2\sigma^2_{N,j}}, \qquad j \in \mathrm{n}(i) \tag{7.28}$$

while for the a priori SNR of the sample A_i we use

$$\hat{\xi}_i^l = \frac{R_{\mathrm{i}}^2}{2\sigma_{N,\mathrm{i}}^2} - 1 \tag{7.29}$$

The above strategy for the estimation of the parameters that are involved in eqs. 7.25 and 7.26 allows us to write the ACMRF estimator in a very compact form and provides some further intuition on its behaviour. We begin by noting that according to eq. 7.28, $\sqrt{\rho_{ij}\hat{\xi}_j^l}\hat{A}_j = R_i\hat{\xi}_j^l$. Using this last result, substitution of the expressions for θ_i and b_{ij} (eqs. 7.25 and 7.26) in the equation for ζ_1 (eq. 7.24) will yield after a simple algebraic manipulation

$$\zeta_1 = \frac{w_{ii}\hat{\xi}_i^l\, R_i + \sum_{j\in \mathrm{n}(i)} w_{ij}\hat{\xi}_j^l\, R_i}{2\left(w_{ii}\hat{\xi}_i^l + \sum_{j\in \mathrm{n}(i)} w_{ij}\hat{\xi}_j^l + a/2\right)} \tag{7.30}$$

If we denote by $\hat{\xi}_i^g$ a 'global' estimate of the a priori SNR at sample i, which we define as

$$\hat{\xi}_i^g \equiv w_{ii}\hat{\xi}_i^l + \sum_{j\in \mathrm{n}(i)} w_{ij}\hat{\xi}_j^l \tag{7.31}$$

then ζ_1 can be further simplified to

$$\zeta_1 = \frac{\hat{\xi}_i^g\, R_i}{2\left(\hat{\xi}_i^g + a/2\right)} \tag{7.32}$$

Following the same procedure, ζ_2 can be reduced to

$$\zeta_2 = (1.5 - a)\frac{\sigma_{N,i}^2 w_{ii}\hat{\xi}_i^l}{\hat{\xi}_i^g + a/2} \tag{7.33}$$

The ACMRF estimator can therefore be summarised as

$$\hat{A}_i = \zeta_1 + \sqrt{\zeta_1^2 - \zeta_2} \tag{7.34}$$

where

$$\zeta_1 = \frac{\hat{\xi}_i^g\, R_i}{2\left(\hat{\xi}_i^g + a/2\right)} \qquad\qquad \zeta_2 = (1.5 - a)\frac{\sigma_{N,i}^2 w_{ii}\hat{\xi}_i^l}{\hat{\xi}_i^g + a/2}$$

Recall from §3.3.1 eq. 3.27 that the MP1C estimator was given by

$$\hat{A}_i = \zeta_1 + \sqrt{\zeta_1^2 - \zeta_2} \tag{7.35}$$

where

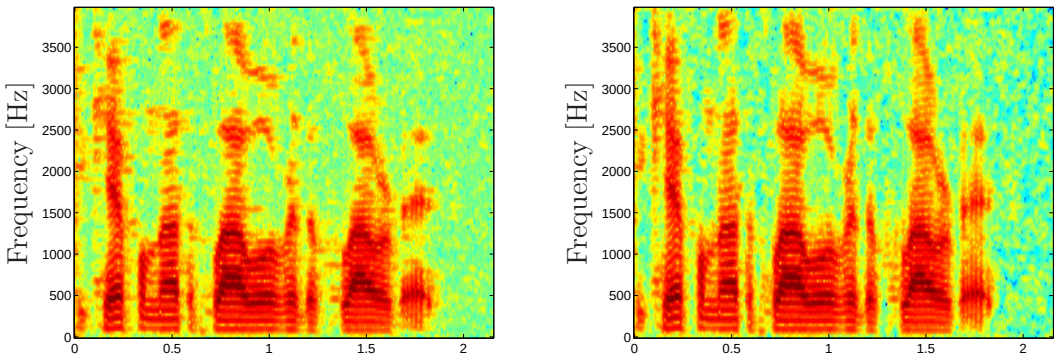
$$\zeta_1 = \frac{\xi_i \, R_i}{2\,(\xi_i + a/2)} \qquad\qquad \zeta_2 = (1.5 - a)\frac{\sigma_{N,i}^2 \xi_i}{\xi_i + a/2}$$

The term ξ_i in the above equations denotes the a priori SNR, which is calculated with the DD method. Therefore, apart from the difference between the numerators in the definition of the ζ_2 the ACMRF estimator can be viewed as the MP1C with a time frequency extended method for the estimation of the a priori SNR.

7.3.4 Results

In our implementation of the ACMRF estimator the values for the weights w_{ij} we have used are $w_{i\mathrm{S}} = 0.48$, $w_{i\mathrm{W}} = 0.49$, $w_{i\mathrm{N}} = 0.01$, $w_{i\mathrm{E}} = 0.01$ and $w_{ii} = 0.01$. Significantly larger weights are placed on the south and west neighbours because the local a priori SNR's $\hat{\xi}_{\mathrm{S}}$ and $\hat{\xi}_{\mathrm{W}}$ are estimated from A_{S} and A_{W}, which were already estimated with the ACMRF algorithm before the estimation of A_i and therefore are more reliable. Conversely, the remaining local a priori SNR's are estimated with a less reliable, but computationally more efficient, power spectral subtraction approach (i.e. $\hat{\xi}_{\mathrm{N}} = \left(R_{\mathrm{N}}^2/2\sigma_{N,\mathrm{N}}^2 - 1\right)^{0.5}$), which typically results in estimates of increased variance. Experiments have shown that increasing the values of $w_{i\mathrm{N}}$, $w_{i\mathrm{E}}$ and w_{ii} at the expense of $w_{i\mathrm{S}}$ and $w_{i\mathrm{W}}$ typically results in musical residual noise. Finally, a lower limit of -25 dB is placed at the global a priori SNR $\hat{\xi}_i^g$, because it contributes to the uniform character of the residual noise. The simulation setup for the evaluation of the ACMRF algorithm is the same to that described in §7.2.4, with the exception of the DD method, which is not required by this algorithm and therefore is not used.

We initially consider the performance of the ACMRF algorithm in comparison to the MP1C algorithm of chapter 3. Figure 7.5(a) shows a speech utterance processed with the MP1C algorithm and $a = 2$, while figure 7.5(b) shows the same utterance processed with the ACMRF algorithm also using $a = 2$. Observe that the ACMRF algorithm provides a significant improvement in the preservation of speech over the



(a) MP1C (b) ACMRF

Figure 7.5: Speech enhanced with the MP1C and ACMRF algorithms using $a = 2$ for both.

MP1C algorithm. Furthermore, this improvement does not come at the expense of an increase in the residual noise level, which as figure 7.5 shows is approximately the same for both algorithms. Informal listening tests confirm that the residual noise of the ACMRF algorithm has a uniform character, although the residual noise of the MP1C is more similar in quality to the original noise. Finally, the time and particularly the frequency coupling of the ACMRF algorithm decreases drastically the number of isolated spectral peaks. This is a significant advantage from a perceptual point of view, because isolated spectral peaks in the vicinity of the main corpus of the speech energy were judged as being quite harmful during the subjective test performed in §5.4.

We now proceed to investigate the effect of the parameter a on the enhanced speech. For $a < 1.5$ the ACMRF estimator is again not well defined, as the discriminant appearing in eq. 7.34 can be negative. The strategy employed is to use the resulting estimate only when the discriminant is positive and suppress the noisy sample by 50 dB otherwise. An example of an utterance enhanced with the ACMRF algorithm and $a = 1$ is shown in figure 7.6(a). Such low values of a are successful in restoring a large number of speech spectral components, particularly in the voiced segments of the utterance. However, the rather strong time frequency coupling in combination with the hard threshold of this algorithm, generates a number of isolated spectral peaks, which are perceived as musical tones and speech distortion.

The above problem is alleviated for $a > 1.5$, because the the estimator is well defined

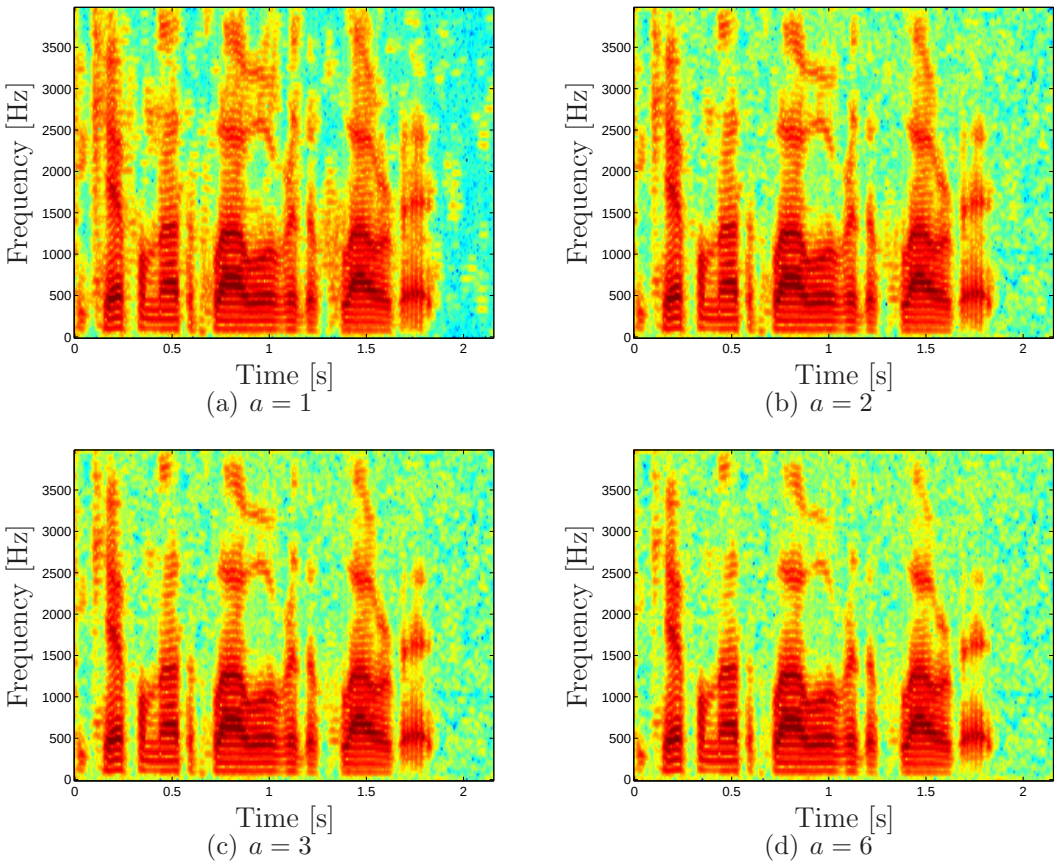
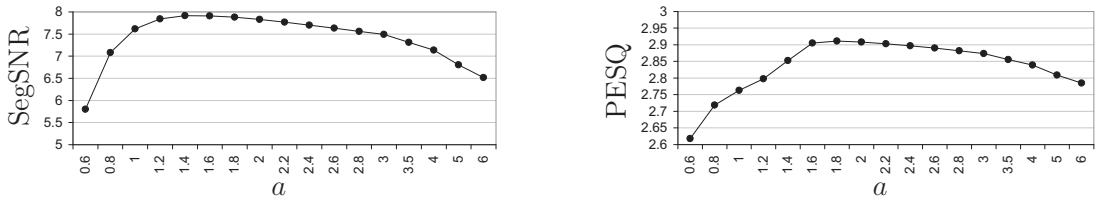


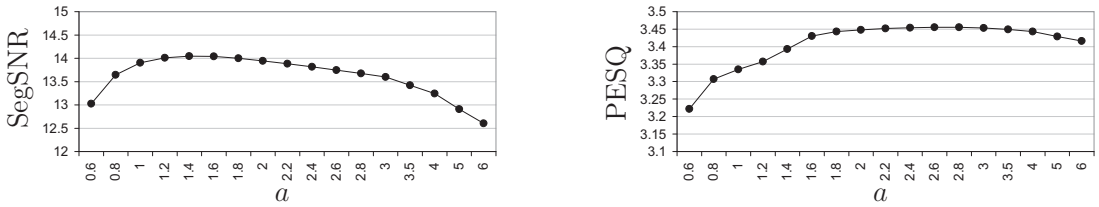
Figure 7.6: Speech enhanced with the ACMRF algorithm and different values of a.

for all the values of its input parameters. This eliminates the isolated spectral peaks, while the residual noise has uniform character and the speech distortions are minimised. An example of the ACMRF algorithm with $a = 2$ is shown in figure 7.6(b). A further increase of a does not alter the quality of noise, but results in the loss of some of the weaker speech spectral components. This effect is rather mild for a as large as 3 (figure 7.6(c)), but becomes more severe as a increases further (figure 7.6(d)). The reason behind the underestimation of the weaker speech spectral components with increasing values of a, is mainly the $a/2$ factor in the denominator of ζ_1 in eq. 7.34, which results in a decrease of the estimates $\hat{A}_i$ as a increases.

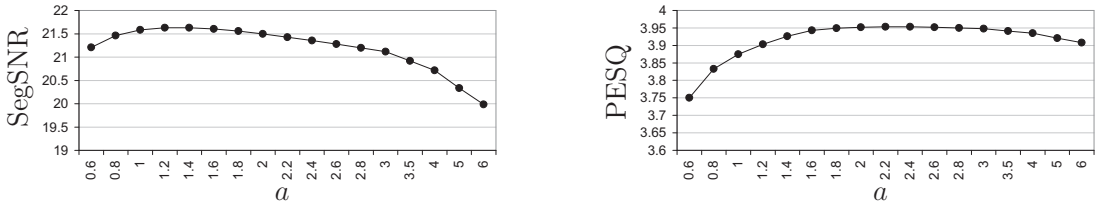
Figures 7.7, 7.8 show the scores in the objective measures of the ACMRF algorithm for different input SegSNR levels and noise types. The results were obtained with the speech database described in §7.2.4. For all the presented cases, the SegSNR exhibits a maximum around $a = 1.4$ and drops monotonically for larger a's. The PESQ on the other hand, reaches its maximum for $1.8 < a < 3$, but also maintains a



(a) Input SegSNR = 0 dB, Input PESQ = 2.11



(b) Input SegSNR = 10 dB, Input PESQ = 2.80

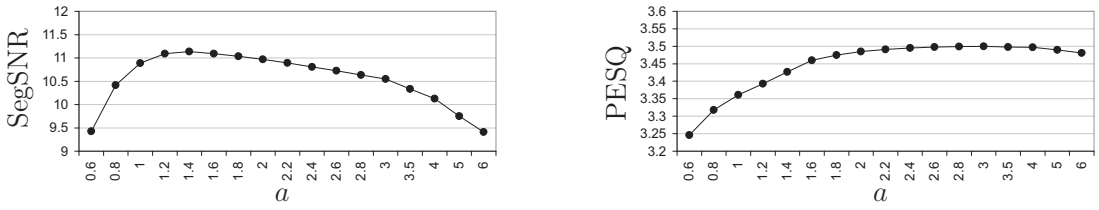


(c) Input SegSNR = 20 dB, Input PESQ = 3.46

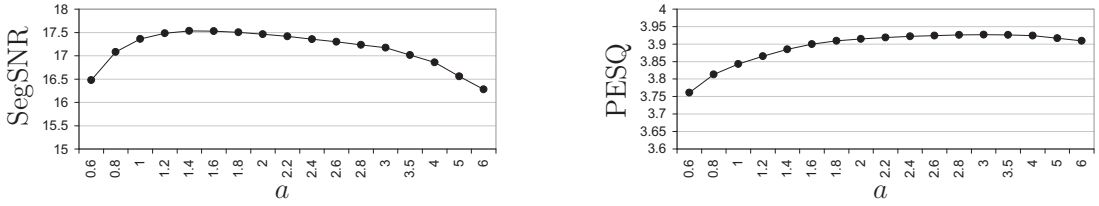
Figure 7.7: Performance of the ACMRF algorithm as function of the parameter a for white Gaussian noise.

fairly constant value within this range. The values of PESQ signify that for $a > 1.8$ the algorithm results in uniform residual noise and minimal speech distortions, while the loss of speech spectral components is not perceptually significant for a as large as 3. This range of a's is considered as the most useful for this algorithm.

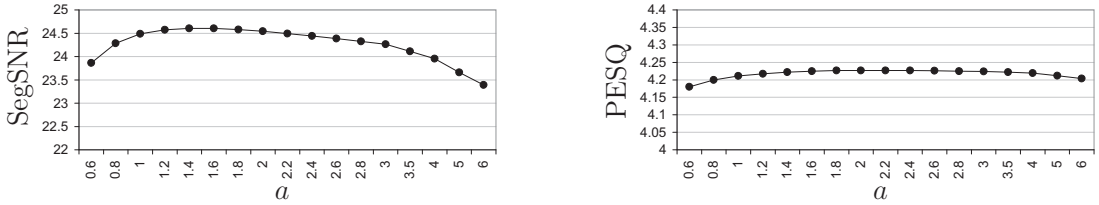
Although subjective tests, such as those in §5.4, could be performed in order to identify an optimum value for a, we believe it is not as necessary for the ACMRF algorithm as it was for the algorithms of chapter 3. Recall that for the algorithms in chapter 3 the parameter a represented a trade off between the level of the residual noise and its musical character. For the ACMRF algorithm however, there seems to be an optimum range between 1.8 and 3. Smaller values result in musical noise and distortion, whereas higher values result in an underestimation of speech components, and both extremes seem to lack any obvious advantage. Furthermore, the objective measures scores do not vary significantly (esp. the PESQ) for $1.8 < a < 3$, which



(a) Input SegSNR = 0 dB, Input PESQ = 2.11



(b) Input SegSNR = 10 dB, Input PESQ = 2.80



(c) Input SegSNR = 20 dB, Input PESQ = 3.46

Figure 7.8: Performance of the ACMRF algorithm as function of the parameter a for car noise.

indicates that the quality of speech, from a perceptual point of view at least, remains fairly constant.

Finally, we provide a comparison between the the MP1G, MS1C and ACMRF algorithms, using for the first two the values of a obtained via the subjective experiments. For the ACMRF algorithm, we use $a = 2$. Table 7.6 shows the scores of the objective measures for the three algorithms and the three different input SegSNR's for white noise. The respective results for car noise are shown in table 7.7. Both tables reveal that the ACMRF algorithm yields consistently higher scores than the MP1C and MS1C algorithms.

Figure 7.9 shows spectrograms of an utterance enhanced with the three algorithms. A conclusion drawn from chapter 5 was that the although the MAP algorithms result in lower levels of residual noise, the MMSE algorithms are more successful in the preservation of the speech spectral components. Figure 7.9 reveals that the ACMRF

Input SegSNR		0 dB		10 dB		20 dB	
		SegSNR	PESQ	SegSNR	PESQ	SegSNR	PESQ
MP1G	$a = 2.6$	7.19	2.74	12.96	3.24	20.40	3.78
MS1C	$a = 1.4$	6.71	2.81	13.03	3.33	20.69	3.82
ACMRF	$a = 2$	7.83	2.91	13.95	3.45	21.50	3.95

Table 7.6: Comparative results for the MP1G, MS1C and ACMRF algorithms for white Gaussian noise.

Input SegSNR		0 dB		10 dB		20 dB	
		SegSNR	PESQ	SegSNR	PESQ	SegSNR	PESQ
MP1G	$a = 2.6$	10.62	3.36	16.79	3.80	24.02	4.18
MS1C	$a = 1.4$	9.72	3.37	16.61	3.82	24.13	4.20
ACMRF	$a = 2$	10.97	3.49	17.46	3.92	24.54	4.23

Table 7.7: Comparative results for the MP1G, MS1C and ACMRF algorithms for car noise.

algorithm combines the advantages of both MAP and MMSE algorithms. It is able to provide residual noise levels similar to that of the MAP, while the preservation of the speech spectral components surpasses that of the MMSE.

7.3.5 Discussion - Motivation

In the previous section we have seen that the ACMRF algorithm is able to restore the weaker speech spectral components while keeping the level of the residual noise low. This behaviour was not attainable from the CMRF algorithm, which used fixed weights between the neighbours. Nevertheless, unlike the CMRF algorithm, the ACMRF has a theoretical weakness: it is not possible to define a valid joint probability density function because the symmetry condition $b_{ij}/\theta_i = b_{ji}/\theta_j$ is not satisfied. We can see this by substituting the expressions for θ_i and b_{ij} from eqs. 7.25, 7.26 in the symmetry condition equation, which yields

$$\frac{w_{ij}\sqrt{\rho_{ij}\xi_j^l}}{2\sigma_{N,i}^2 w_{ii}\xi_i^l} \neq \frac{w_{ii}\sqrt{\rho_{ji}\xi_i^l}}{2\sigma_{N,j}^2 w_{ij}\xi_j^l} \tag{7.36}$$

The ACMRF therefore cannot be considered as an MRF algorithm in the strict sense, because it yields no valid joint probability density function. Instead the ACMRF could be seen as an MRF-based or MRF-inspired algorithm. However, it is probably

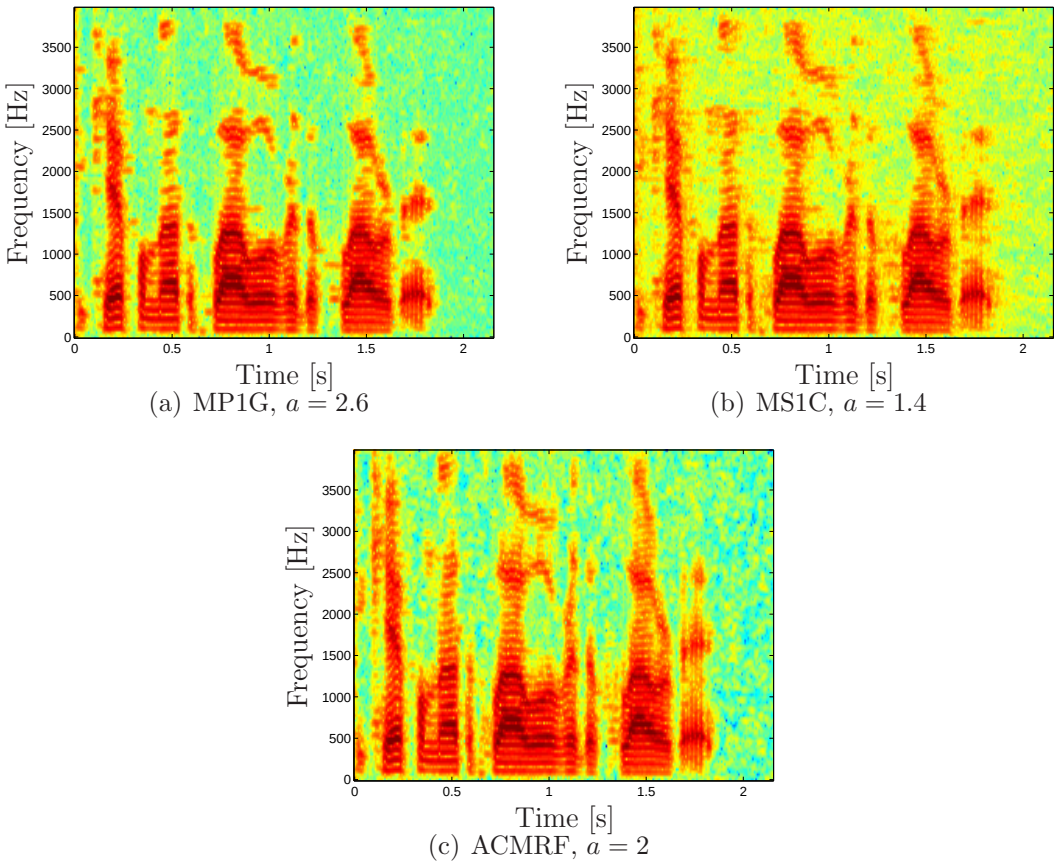


Figure 7.9: Speech enhanced with the MP1G, MS1C and ACMRF algorithms.

fair to say that the above remark has a more theoretical than practical significance. After all, the ICM estimation, on which the ACMRF is based, does not require the existence of a valid joint density, as the global optimisation methods do [11].

In the presentation of the update equations for θ_i and b_{ij} (eqs. 7.25, 7.26), we mentioned that the update for θ_i was essentially a ratio between local SNR's, proportional to the local SNR of the i^{th} sample. The process that led to the derivation of the update equation for b_{ij} was delayed however until the end of the chapter, because it is slightly more involved, but we believe that it provides interesting insights on the application of MRF's to speech enhancement.

After the failure of the CMRF algorithm with the fixed weights to restore the speech spectral components without increasing the residual noise level, we considered altering the influence of the neighbours depending on their local SNR. The rationale was that a sample with high SNR would typically correspond to speech, therefore it should contribute to the final estimate, while a low SNR sample would mainly

contain noise, so it should be excluded. The weights first considered were

$$\frac{\theta_i}{2\sigma^2_{N,i}} = \frac{w_{ii}\xi^l_i}{\sum_{m\in \mathrm{n}(i)} w_{im}\xi^l_m + a/2} \tag{7.37}$$

and

$$b_{ij} = \frac{w_{ij}\xi^l_j}{\sum_{m\in \mathrm{n}(i)} w_{im}\xi^l_m + a/2} \tag{7.38}$$

According to these weights, the influence of a neighbour would be proportional to its SNR and in the limit if $\xi^l_i >>$ then $\hat{A}_i = R_i$, while if $\xi^l_j >>$ then $\hat{A}_i = A_j$.

Under this scenario however, the role of the evidence, provided by the noisy speech R_i acts 'competitively' with the influence of the neighbours. That is, if the weights of the neighbours w_{ij} increase, the weight of the evidence R_i decreases, assuming that $\sum_{j\in \mathrm{n}(i)} w_{ij} + w_{ii} = 1$. Additionally, we have found that using the value of a neighbour A_j as an estimate for A_i, (e.g. if A_j has a high a priori SNR), generates annoying speech artifacts. In an image processing scenario, where the MRF's have been extensively used, substitution of a pixel's value with that of its neighbour's might be desirable, assuming that both pixels represent the same color. This approach to the restoration of the speech STFT amplitudes however, was found to generate significant distortions.

The proposed parameters (eqs. 7.25, 7.26) on the other hand, avoid the direct substitution of the neighbour values. Instead, as it can be seen from the form of the estimator shown in eq. 7.34, the spectral amplitude of the neighbours indicates the amount of the suppression that has to be applied to the noisy sample R_i, by means of the a 'global' estimate of the a priori SNR $\hat{\xi}^g_i$. This allows the restoration of weak spectral samples that lie in a neighbourhood of samples with large amplitude, while it avoids the artifacts generated by the direct substitution of the neighbours' spectral amplitude values.

7.4 Summary

In this chapter we proposed and investigated the application of MRF's to the problem of enhancing speech that is corrupted with broadband noise. This study was

triggered by our desire to incorporate the time and frequency dependencies of speech signals into the estimation model.

We first developed an algorithm based on Gaussian MRF priors. This algorithm resulted in an improvement in the preservation of speech spectral components by coupling the the STFT samples both in time and in frequency. The algorithm's major drawback however, was that the estimator was not well defined for all the values of its input parameter, due to an approximation in its derivation that allowed the estimator to be expressed in a closed form. This resulted in the amplification of some isolated spectral peaks, which were then perceived as musical noise.

In order to overcome this problem, we introduced the Chi MRF priors, proving as well that they result in valid MRF schemes. A MAP estimator based on the Chi MRF priors was then derived. When the latter estimator was applied with fixed weights between the neighbouring samples, the time frequency coupling enhanced the weaker speech spectral components, but the level of the residual noise also increased. This was attributed to the fact that the fixed neighbours' weights were not designed to differentiate between the samples that contained speech plus noise or noise only, thus increasing the level of both.

An adaptive scheme for the estimation of the neighbours' weights was finally devised, which was capable of performing the above differentiation. The result was an algorithm which enhanced the spectral components that belonged to speech, while keeping the level of the residual noise low. The proposed adaptive scheme was shown to combine the low residual noise levels of the MAP algorithms of chapter 3, with the ability to surpass the MMSE algorithms of the same chapter in restoring the spectral components that belong to speech.

Chapter 8

Conclusion

This thesis considers the problem of enhancing speech that has been corrupted with additive and uncorrelated noise. The problem of speech enhancement was formulated as an estimation problem in the STFT domain, according to which, an optimal, in some sense, estimate of the clean speech STFT was sought, when only the noisy speech STFT was observed. Given the above formulation of the problem and with a number of tools from the Bayesian machinery at our disposal, we have made several novel contributions in the field of speech enhancement, all of which are summarised in the next section, along with the conclusions that have been reached during the course of this work. An outline of ideas that build on the methods and concepts developed in this thesis and can potentially produce fruitful research results is presented in §8.2.

8.1 Summary - conclusions

The research that has been carried out in this project can be divided in three main parts. In the first part (chapters 3 - 5) a framework of Bayesian algorithms for speech enhancement was proposed and studied, which consists of the generalisation of existing algorithms and the introduction of novel ones. The second part (chapter 6) is concerned with the development of algorithms that can estimate the power of time varying noises and can be used with those algorithms that estimate the clean speech STFT coefficients. Finally, the third part (chapter 7) is a study on the application of Markov Random Fields to speech enhancement, where the incorporation of

the time and frequency dependencies of speech in the estimation model was sought.

Apart from offering an opportunity to compare directly several successful speech enhancement algorithms from the literature, the compilation of the framework of algorithms in chapter 3 provides insight on the effect of several components of a Bayesian STFT estimation speech enhancement algorithm to the quality of the enhanced speech. The components studied were the estimated feature (Re and Im parts and the amplitude of the STFT), the employed estimator (MMSE and MAP) and the shape and type of the speech prior (Chi, Gamma and Lognormal).

As mentioned previously, some members of the family of algorithms presented in chapter 3 are generalisations of existing algorithms, while others are proposed for the first time in this work. Specifically, the MS1C algorithm is a generalisation of the Ephraim-Malah algorithm [31] as the Wiener filter [63, 72] is a special case of the MS2C. The algorithms proposed by Martin [72] are special cases of the MS2G, and the MP1C and MP1G algorithms are generalisations of the algorithms found in [24, 66, 99]. By *generalisation* here we mean that the algorithms we propose yield the *special case* algorithms mentioned above for a particular value of the shape parameter that is incorporated in their priors. On the other hand, the DFT MAP algorithms (MP2C, MP2G) are introduced for the first time in this work, and so are the MS1G and the algorithms that use the Lognormal priors (MP1L, MS1L).

The analysis of chapter 5 showed that for the MAP algorithms, the choice of the estimated feature (Re and Im parts or amplitude) had a rather small effect in the quality of the resulting speech. The amplitude MAP algorithms however, were marginally better in the preservation of speech, which might give them an edge over their DFT counterparts. Additionally, as with all the amplitude estimation algorithms, their computational load is smaller because only the amplitude needs to be estimated, while the DFT algorithms require the estimation of both the Re and Im parts.

On the other hand, the selection of the estimated feature played an important role for the MMSE algorithms. The residual noise of the MMSE DFT algorithms had a musical character for all the values of the priors' shape parameter a, which can hinder their employment in audio speech enhancement applications. This problem

was not apparent in the amplitude MMSE algorithms, because appropriate values of a resulted in uniform residual noise.

The type of the employed estimator was very influential in the quality of the enhanced speech. The algorithms that employed the MAP estimator resulted in lower noise levels while they also had a lower computational load. On the other hand, the MMSE based algorithms were more successful in the preservation of speech and generally achieved higher scores in the objective measures.

An interesting observation that emerged from the study of the different priors, was that an appropriate tuning of the priors' shape parameter a could yield speech of very similar quality for all the three families of priors. A possible reason for the similar performances achieved with the three different priors is the flexibility in their shape that is provided by the shape parameter a. Nevertheless, there were some differences in the quality of the resulting speech depending on the employed prior, which are summarised in the following: according to figures 4.5, 4.6, we classify the three priors with respect to the length of their tails as Chi (shorter tails), Gamma and Lognormal (longer tails). The combination of a short tailed prior with a MAP estimator results in the preservation of a few extra speech spectral components, but a long tailed prior results in slightly less distorted speech. A long tailed prior in combination with an MMSE estimator results in a somewhat better preservation of speech, especially at its onset, but a shorter tailed prior results in more uniform residual noise.

In chapter 4 we tried to extract optimal values for both the shape and the scale parameters a and θ of the priors. To realise this goal, two methods were employed: the first consisted of fitting the priors to a large number of clean speech data, via the minimisation of the KL divergence, while the second was based on adaptive estimation of the parameters. The adaptive method was preferred for the estimation of the scale parameter, because using fixed values of θ resulted in high levels of musical residual noise. In accordance with the practice followed in the relevant literature (e.g. [31,72,99]), the scale parameter was estimated from the a priori SNR, which was in turn calculated with the DD method. The adaptive estimation of θ excludes the use of long term speech data for the estimation of the shape parameter a, because fitting the priors to such data assumes a fixed value for θ. A method

for estimating a from narrow a priori SNR intervals, which is compatible with the adaptive estimation model of θ, was also implemented, but failed to produce consistent results for data selected from different a priori SNR intervals. Additionally, the method for the adaptive estimation of a, which was based on moment matching showed limited success.

In view of the shortcomings of the above methods, the approach we followed was to evaluate the performance of the proposed algorithms as a function of the priors' shape parameter a and reach an a posteriori decision for their optimal values, based on the performance results. The analysis of chapter 5 revealed that the shape parameter a essentially controls a trade off between the musical character of the residual noise and its level. Small values of a, which correspond to priors with large concentration around zero and heavy tails, result in good preservation of speech but the residual noise has a strong musical character. Large values of a, which correspond to flatter priors, result in an increase in the level of the residual noise and to an underestimation of the speech components, primarily for the MAP algorithms, but their significant benefit is that the residual noise has a uniform character.

In order to identify an optimal value for the shape parameter a we carried out formal subjective listening tests. During these tests a panel of listeners was asked to tune the shape parameter a so that the quality of the enhanced speech is maximised. An interesting conclusion that stemmed from the subjective tests was that the selected values of a were significantly different from those which maximised the scores of the objective measures. This may be an indication that there might be further room for improving the objective speech quality measures that attempt to predict the subjective quality of speech.

In chapter 6, we developed methods for estimating the power of time varying noise, which is an essential component of every single channel speech enhancement scheme. An algorithm based on Gaussian Mixture Models of noise was developed, capable of modelling very accurately the distribution of time varying noise STFT coefficients. The results however showed that a noise model based on a single Gaussian distribution is preferable and more simple to implement, as long as there is an algorithm that can effectively track the variations of the time varying noise power. Such an algorithm was also proposed in chapter 6, which was based on an observation about

the distribution of the noisy speech spectral amplitude coefficients that had received little attention in the literature. The main benefit of this algorithm was the quick adaptation of the estimates in the event of an increase in the noise power. Its main disadvantage was its tendency to overestimate the noise power in periods of prolonged speech activity. Nevertheless, its overall performance was comparable with the performance of state of the art noise estimation methods such as the minimum statistics method proposed in [71].

In the final part of this thesis, we employed tools from the theory of Markov Random Fields, in order to create models that account for the time and frequency dependencies of speech signals. We first developed an algorithm based on Gaussian MRF priors, which, despite its success in introducing the time and frequency dependencies in the estimation model, was not well defined for all the values of its input parameters and resulted in speech suffering from musical noise. In order to overcome the deficiencies of this algorithm we proposed a novel type of MRF, which we termed Chi MRF, proving also its validity as an MRF model. The major outcome was the development of an adaptive algorithm based on Chi MRF's, which combined low levels of uniform residual noise - the strong point of the MAP algorithms of chapter 3 - with the ability to surpass the MMSE algorithms of the same chapter in the restoration of the weaker speech spectral components.

8.2 Further work

The analysis of chapter 5 showed that algorithms which used different priors but the same combination of estimator and estimated feature (e.g. MP1C and MP1G) resulted in speech of very similar quality when the priors' shape parameters were tuned appropriately (i.e. tuning a so that the levels of residual noise for two different algorithms are equalised). This can be an indication that there is probably little margin for improving an algorithm's performance by experimenting with different density functions that model individually the speech spectral samples. On the other hand, the two employed estimators (MMSE and MAP) resulted in speech of significantly different quality, which suggests that experimentation with alternative estimators can yield interesting results. An example could be the combination of one of the studied priors (e.g. Chi) with the Log spectral estimator proposed in [32].

Additionally, the combination algorithms proposed in chapter 3 with the method of Ephraim and Malah [31] that takes into account the uncertainty of the speech presence, is also worth considering.

Given the importance of the DD method to the enhanced speech quality, and the similarity of the ACMRF algorithm of chapter 7 with the MP1C that utilises a time frequency extended DD estimator of the a priori SNR, we may conclude that research into improving the method for the estimation of the a priori SNR has a high probability of producing successful speech enhancement schemes. Steps towards this direction have appeared recently in the literature [21, 46].

The discrepancy between the values of a extracted via our formal listening tests and the values of a that maximised the scores of the objective measures also indicate that there is margin for improving the algorithms that objectively evaluate the speech quality. In particular, the PESQ measure appeared to be relatively insensitive to the musical residual noise, which was judged as annoying by the participants in our test. Addressing the above issue could provide a more robust evaluation measure and possibly reduce the need to perform formal subjective tests.

The noise estimation algorithm we proposed in chapter 6 presented encouraging results but was marred by the overestimation of noise during periods of prolonged speech activity. As the review of the noise estimation methods showed in §6.1, a current trend in the field is the merging of elements from different methods, (e.g. averaging and minimum statistics §6.1.3). Elements of these methods could also be combined with the principles of the method we proposed in §6.3, for its further improvement.

Additionally, the interaction between the noise estimation and the speech estimation modules of a speech enhancement scheme could be an interesting field of research. An objective of this research could be the development of a speech enhancement scheme, in which it is not only the speech estimation algorithm that uses the estimates of the noise module, but there is instead a closer interaction of the two modules for improving the performance of both.

The application of Markov Random Fields to speech enhancement is a novel idea that has produced very good results so far. We believe that the MRF's represent

a powerful tool in the development of speech enhancement algorithms and that this thesis has only scratched the surface of their potential. Their main strength lies in that they provide a framework for encapsulating the time and frequency dependencies of speech in the estimation model. In the following, we mention some of the directions into which the relevant research could expand.

The parameters of the ACMRF algorithm were chosen empirically based on the requirement that the speech spectral components are enhanced, while the residual noise level is kept to a minimum. Alternative parameterisations could also yield interesting results, while they could be selected either empirically, as the ones we proposed, or based on standard statistical procedures, as the Maximum Likelihood method presented in [80].

The ACMRF algorithm, which is based on Iterated Conditional Modes, performs only a single iteration. Algorithms with multiple iterations could also be developed, in an effort to reduce further the level of the residual noise and improve the preservation of the speech spectral components. The exploration of alternative neighbourhood structures could also be another extension of the presented work. For example, the influence of STFT points that are further apart in time and in frequency could be incorporated in the existing models in a straightforward way.

As it was shown in appendix D eq. D.5 the joint Gaussian MRF density can be written as

$$p(x) \propto \exp\left[-\sum_{i\in Q} b'_i\, x_i^2 - \sum_{\{i,j\}\in C} b'_{ij}\,(x_i - x_j)^2\right]$$

where b'_i and b'_{ij} is a shorthand notation for the parameters of the prior as they appear in eq. D.5. Bouman and Sauer [14] proposed a generalised Gaussian MRF prior for image processing problems, which has the form

$$p(x) \propto \exp\left[-\sum_{i\in Q} b'_i\, x_i^{\beta} - \sum_{\{i,j\}\in C} b'_{ij}\,(x_i - x_j)^{\beta}\right]$$

where β is a real number in the interval [1,2]. The effect of these generalised priors on speech enhancement algorithms could also be investigated, as it has happened with the investigation of spectral subtraction of arbitrary powers of the speech spectrum

[9] or the MMSE estimators of an arbitrary power of the speech spectral amplitude [100]. Furthermore, research into the development of MRF's based on alternative density functions (e.g. Lognormal) is also possible.

Cohen [21] proposed a model in which dependencies exist between the speech spectral variances, while the speech spectral amplitude samples are independent, given the value of their variance. In the same spirit, the speech spectral variances could be modelled with an MRF, allowing for a variety of estimators of the speech spectral amplitude to be applied (e.g. MMSE, LogMMSE), while preserving the time and frequency dependencies of the model.

Appendix A

Derivation of the estimators

A.1 Derivation of the amplitude posterior density

According to eq. 2.25 the likelihood $p(\mathbf{X}|\mathbf{S})$ can be written as

$$p(\mathbf{X}|\mathbf{S}) = p_N(\mathbf{X} - \mathbf{S}) \tag{A.1}$$

where p_N is the pdf of the noise STFT coefficients. Assuming that these are Gaussian and independent random variables with zero mean and variance σ_N^2 eq. A.1 can be written as:

$$p(\mathbf{X}|\mathbf{S}) \equiv p(X_{\mathrm{Re}}, X_{\mathrm{Im}}|S_{\mathrm{Re}}, S_{\mathrm{Im}}) = \frac{1}{2\pi\sigma_N^2} \exp\left[-\frac{(X_{\mathrm{Re}} - S_{\mathrm{Re}})^2 + (X_{\mathrm{Im}} - S_{\mathrm{Im}})^2}{2\sigma_N^2}\right] \tag{A.2}$$

where X_{Re} and X_{Im} denote the Re and Im parts of $\mathbf{X}$ and similarly for $\mathbf{S}$.

Our goal is to find $p(R, \psi|A, \phi)$ when we know $p(\mathbf{X}|\mathbf{S})$. If we define by $D_{R\psi}$ the slice of a circle of radius R_0 and angle ψ_0 centered at zero on the plane X_{Re}, X_{Im}, the probability mass that it encloses can be written as

$$P_{R,\psi|A,\phi}(R_0, \psi_0|A, \phi) = \iint_{D_{R\psi}} p(X_{\mathrm{Re}}, X_{\mathrm{Im}}|S_{\mathrm{Re}}, S_{\mathrm{Im}})\, dX_{\mathrm{Re}}\, dX_{\mathrm{Im}} \tag{A.3}$$

where $P_{R,\psi|A,\phi}(R_0, \psi_0|A, \phi)$ is the probability *distribution* function of R and ψ given A and ϕ, or in other words, the probability that $R \leq R_0$ and $\psi \leq \psi_0$ given A and ϕ. If we change the Cartesian to polar coordinates in the integral in eq. A.3 (i.e.

$X_{\mathrm{Re}} = r\cos\omega$, $X_{\mathrm{Im}} = r\sin\omega$ and $dX_{\mathrm{Re}}\, dX_{\mathrm{Im}} = rdrd\omega$) and express $S_{\mathrm{Re}}, S_{\mathrm{Im}}$ in their polar form A, ϕ we get:

$$P_{R,\psi|A,\phi}(R_0, \psi_0|A, \phi) = \int_0^{R_0}\int_0^{\psi_0} p(r, \omega|A, \phi) r\ dr\, d\omega \tag{A.4}$$

Substituting the expression for $p(X_{\mathrm{Re}}, X_{\mathrm{Im}}|S_{\mathrm{Re}}, S_{\mathrm{Im}})$ from eq. A.2 we have:

$$\begin{aligned} &P_{R,\psi|A,\phi}(R_0, \psi_0|A, \phi) = \\ &\frac{1}{2\pi\sigma_N^2} \int_0^{R_0}\int_0^{\psi_0} \exp\left[-\frac{(r\cos\omega - A\cos\phi)^2 + (r\sin\omega - r\sin\phi)^2}{2\sigma_N^2}\right] r\ dr\, d\omega = \\ &\frac{1}{2\pi\sigma_N^2} \int_0^{R_0}\int_0^{\psi_0} \exp\left[-\frac{r^2 + A^2 - 2rA\cos(\omega - \phi)}{2\sigma_N^2}\right] r\ dr\, d\omega \end{aligned} \tag{A.5}$$

The probability density function of R and ψ given A and ϕ is easily obtained by differentiating the distribution function with respect to R_0 and ψ_0.

$$\begin{aligned} p_{R,\psi|A,\phi}(R_0, \psi_0|A, \phi) = \frac{\partial^2}{\partial R_0\, \partial\psi_0}\ P_{R,\psi|A,\phi}(R_0, \psi_0|A, \phi) = \\ \frac{R_0}{2\pi\sigma_N^2}\ \exp\left[-\frac{R_0^2 + A^2 - 2R_0A\cos(\psi_0 - \phi)}{2\sigma_N^2}\right] \end{aligned} \tag{A.6}$$

Finally, by denoting R_0 and ψ_0 with R and ψ we have:

$$p(R, \psi|A, \phi) = \frac{R}{2\pi\sigma_N^2}\ \exp\left[-\frac{R^2 + A^2 - 2RA\cos(\psi - \phi)}{2\sigma_N^2}\right] \tag{A.7}$$

Integration of the phases ϕ and ψ can also yield an expression for the $p(R|A)$.

$$\int_0^{2\pi} p(R, \psi|A, \phi)\, d\psi = \frac{R}{\sigma_N^2}\ \exp\left[-\frac{R^2 + A^2}{2\sigma_N^2}\right] \frac{1}{2\pi}\int_0^{2\pi} \exp\left[\frac{RA\cos(\psi - \phi)}{\sigma_N^2}\right] d\psi \tag{A.8}$$

Using eq. 8.431.5 from [42]

$$I_0\left(\frac{RA}{\sigma_N^2}\right) = \frac{1}{2\pi}\int_0^{2\pi} \exp\left[\frac{RA\cos(\psi - \phi)}{\sigma_N^2}\right] d\phi \tag{A.9}$$

where $I_0(z)$ is the modified Bessel function of the first kind we have

$$p(R|A,\phi) = \frac{R}{\sigma_N^2}\ \exp\left[-\frac{R^2+A^2}{2\sigma_N^2}\right] I_0\left(\frac{RA}{\sigma_N^2}\right) \tag{A.10}$$

With the assumption that the phase ϕ is uniformly distributed (i.e. $p(\phi) = \frac{1}{2\pi}$) we have

$$p(R|A,\phi) = \frac{p(R,\phi|A)}{p(\phi)} = p(R|A) \tag{A.11}$$

since R and ϕ are independent conditioned on A (i.e. $p(R,\phi|A) = p(R|A)p(\phi)$). Therefore

$$p(R|A) = \frac{R}{\sigma_N^2}\ \exp\left[-\frac{R^2+A^2}{2\sigma_N^2}\right] I_0\left(\frac{RA}{\sigma_N^2}\right) \tag{A.12}$$

An approximate expression for the above equation can be found by using the approximation [1] for the Bessel function [73]

$$I_0(z) \sim \mathrm{e}^z/\sqrt{2\pi z} \tag{A.13}$$

The approximate expression for $p(R|A)$ then reads

$$p(R|A) \sim \sqrt{\frac{R}{2\pi\sigma_N^2 A}}\ \exp\left[-\frac{R^2+A^2}{2\sigma_N^2}\right] \exp\left[\frac{RA}{\sigma_N^2}\right] \tag{A.14}$$

A.2 Derivation of the MS2C estimator

Substitution of eqs. 3.7 and 3.5 into eq. 3.8 yields:

$$\hat{S} = \frac{\displaystyle\int_{-\infty}^{\infty} S \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left[-\frac{(X-S)^2}{2\sigma_N^2}\right] \frac{|S|^{a-1}}{\theta^{a/2}\Gamma(a/2)} \exp\left[-\frac{S^2}{\theta}\right] dS}{\displaystyle\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left[-\frac{(X-S)^2}{2\sigma_N^2}\right] \frac{|S|^{a-1}}{\theta^{a/2}\Gamma(a/2)} \exp\left[-\frac{S^2}{\theta}\right] dS} \tag{A.15}$$

[1] The relative error for this approximation is less than 5% for $z > 3$, while the largest discrepancy is found for $z \to 0$, for which value the Bessel function tends to 1, while its approximation tends to infinity.

The numerator can be written as:

$$\begin{aligned}\text{num} &= \int_{-\infty}^{0} S\frac{1}{\sqrt{2\pi\sigma_N^2}}\exp\left[-\frac{(X-S)^2}{2\sigma_N^2}\right]\frac{(-S)^{a-1}}{\theta^{a/2}\Gamma(a/2)}\exp\left[-\frac{S^2}{\theta}\right]dS \\ &+ \int_{0}^{\infty} S\frac{1}{\sqrt{2\pi\sigma_N^2}}\exp\left[-\frac{(X-S)^2}{2\sigma_N^2}\right]\frac{S^{a-1}}{\theta^{a/2}\Gamma(a/2)}\exp\left[-\frac{S^2}{\theta}\right]dS\end{aligned}$$

Making the substitution $S = -S$ in the first integral we have:

$$\begin{aligned}\text{num} &= \int_{0}^{\infty} -S\frac{1}{\sqrt{2\pi\sigma_N^2}}\exp\left[-\frac{(X+S)^2}{2\sigma_N^2}\right]\frac{S^{a-1}}{\theta^{a/2}\Gamma(a/2)}\exp\left[-\frac{S^2}{\theta}\right]dS \\ &+ \int_{0}^{\infty} S\frac{1}{\sqrt{2\pi\sigma_N^2}}\exp\left[-\frac{(X-S)^2}{2\sigma_N^2}\right]\frac{S^{a-1}}{\theta^{a/2}\Gamma(a/2)}\exp\left[-\frac{S^2}{\theta}\right]dS\end{aligned}$$

Expanding the exponentials and taking common factors:

$$\begin{aligned}\text{num} &= \frac{\exp\left[-\frac{X^2}{2\sigma_N^2}\right]}{\sqrt{2\pi\sigma_N^2}\,\theta^{a/2}\,\Gamma(a/2)}\left[-\int_{0}^{\infty} S^a\exp\left[-S^2\left(\frac{1}{2\sigma_N^2}+\frac{1}{\theta}\right)-S\frac{X}{\sigma_N^2}\right]dS\right. \\ &+ \left.\int_{0}^{\infty} S^a\exp\left[-S^2\left(\frac{1}{2\sigma_N^2}+\frac{1}{\theta}\right)-S\frac{X}{\sigma_N^2}\right]dS\right] \qquad \text{(A.16)}\end{aligned}$$

The above integrals can be solved with equation 3.462.1 found in [42], which is stated below.

$$\int_{0}^{\infty} x^{\nu-1}\exp[-\beta x^2-\gamma x]\,dx = (2\beta)^{-\nu/2}\,\Gamma(\nu)\,\exp\left[\frac{\gamma^2}{8\beta}\right]D_{-\nu}\left(\frac{\gamma}{\sqrt{2\beta}}\right) \qquad \text{(A.17)}$$

where $D_\nu(z)$ is the Parabolic Cylinder Function (eq. 9.240, [42]).

Solving the integrals in eq. A.16 according to eq. A.17 we have:

$$\begin{aligned}\text{num} &= \frac{\exp\left[-\frac{X^2}{2\sigma_N^2}\right]}{\sqrt{2\pi\sigma_N^2}\,\theta^{a/2}\,\Gamma(a/2)}\left(\frac{1}{\sigma_N^2}+\frac{2}{\theta}\right)^{-(a+1)/2}\Gamma(a+1)\,\exp\left[\frac{\left(\frac{X}{\sigma_N^2}\right)^2}{8\left(\frac{1}{2\sigma_N^2}+\frac{1}{\theta}\right)}\right] \\ &\left[-D_{-a-1},\left(\frac{\frac{X}{\sigma_N^2}}{\sqrt{\frac{1}{\sigma_N^2}+\frac{2}{\theta}}}\right)+D_{-a-1},\left(\frac{-\frac{X}{\sigma_N^2}}{\sqrt{\frac{1}{\sigma_N^2}+\frac{2}{\theta}}}\right)\right] \qquad \text{(A.18)}\end{aligned}$$

Performing the same steps on the denominator of eq. A.15 we get:

$$\text{den} = \frac{\exp\left[-\frac{X^2}{2\sigma_N^2}\right]}{\sqrt{2\pi\sigma_N^2}\,\theta^{a/2}\,\Gamma(a/2)}\left(\frac{1}{\sigma_N^2}+\frac{2}{\theta}\right)^{-a/2}\Gamma(a)\,\exp\left[\frac{\left(\frac{X}{\sigma_N^2}\right)^2}{8\left(\frac{1}{2\sigma_N^2}+\frac{1}{\theta}\right)}\right]$$
$$\left[D_{-a},\left(\frac{\frac{X}{\sigma_N^2}}{\sqrt{\frac{1}{\sigma_N^2}+\frac{2}{\theta}}}\right)+D_{-a},\left(\frac{-\frac{X}{\sigma_N^2}}{\sqrt{\frac{1}{\sigma_N^2}+\frac{2}{\theta}}}\right)\right] \tag{A.19}$$

Dividing the two above equations we get:

$$\hat{S} = \left(\frac{1}{\sigma_N^2}+\frac{2}{\theta}\right)^{-1/2}\frac{\Gamma(a+1)}{\Gamma(a)}\frac{D_{-a-1}(-\zeta X)-D_{-a-1}(\zeta X)}{D_{-a}(-\zeta X)+D_{-a}(\zeta X)} \tag{A.20}$$

where

$$\zeta = \frac{1/\sigma_N^2}{\sqrt{1/\sigma_N^2+2/\theta}} = \sqrt{\frac{\theta/\sigma_N^2}{\theta+2\sigma_N^2}}$$

Considering that $\Gamma(a+1)/\Gamma(a) = a$ and expressing the first square root of eq. A.20 in terms of ζ we have:

$$\hat{S} = a\sigma_N^2\zeta\,\frac{D_{-a-1}(-\zeta X)-D_{-a-1}(\zeta X)}{D_{-a}(-\zeta X)+D_{-a}(\zeta X)} \qquad \text{where} \quad \zeta = \sqrt{\frac{\theta/\sigma_N^2}{\theta+2\sigma_N^2}} \tag{A.21}$$

A.3 Derivation of the MP2C estimator

The MAP estimate is the value of S which maximises $\ln(p(X|S)p(S))$, where $p(X|S)$ and $p(S)$ are given by 3.5 and 3.7 respectively. We therefore have:

$$\ln(p(X|S)p(S)) = \ln\left[\frac{1}{\sqrt{2\pi\sigma_N^2}}\exp\left[-\frac{(X-S)^2}{2\sigma_N^2}\right]\frac{|S|^{a-1}}{\theta^{a/2}\,\Gamma(a/2)}\exp\left[-\frac{S^2}{\theta}\right]\right]$$

Taking the derivative w.r.t. S we get:

$$\frac{d(\ln(p(X|S)p(S)))}{dS} = \frac{X-S}{\sigma_N^2}+\frac{a-1}{S}-\frac{2S}{\theta} \tag{A.22}$$

Setting the above equation to zero and solving w.r.t S we get:

$$\hat{S} = \zeta \frac{X}{2} + \text{sgn}(X) \left[\left(\zeta \frac{X}{2} \right)^2 + (a-1)\, \sigma_N^2\, \zeta \right]^{1/2} \qquad \text{where} \quad \zeta = \frac{\theta}{\theta + 2\sigma_N^2} \tag{A.23}$$

The above estimator comes from solving a quadratic equation, which can have two solutions. We briefly describe which one is chosen and how the sgn(.) appears in the above equation. The value of S for which the posterior density has its maximum has the same sign as X as it can be seen from the form of $p(X|S)P(S)$. For $a > 1$ the two solutions have different signs, so we chose the one that has the same sign as X. For $a < 1$ both of the solutions have the same sign but one only is a maximum, which is what we are looking for. Following these rules, it turns that the correct sign from the $\pm$ is the one that matches the sign of X.

A.4 Derivation of the MS2G estimator

Substituting in eq. 3.8 the expression for the likelihood (eq. 3.5) and the Gamma prior, which is given by eq. 3.14 we have:

$$\hat{S} = \frac{\displaystyle\int_{-\infty}^{\infty} \frac{S}{\sqrt{2\pi\sigma_N^2}} \exp\left[-\frac{(X-S)^2}{2\sigma_N^2} \right] \frac{|S|^{a-1}}{2\,\theta^a\,\Gamma(a)} \exp\left[-\frac{|S|}{\theta} \right] dS}{\displaystyle\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left[-\frac{(X-S)^2}{2\sigma_N^2} \right] \frac{|S|^{a-1}}{2\,\theta^a\,\Gamma(a)} \exp\left[-\frac{|S|}{\theta} \right] dS} \tag{A.24}$$

The numerator can be written as:

$$\begin{aligned} \text{num} \;&=\; \int_{-\infty}^{0} \frac{S}{\sqrt{2\pi\sigma_N^2}} \exp\left[-\frac{(X-S)^2}{2\sigma_N^2} \right] \frac{(-S)^{a-1}}{2\,\theta^a\,\Gamma(a)} \exp\left[\frac{S}{\theta} \right] dS \\ &+\; \int_{0}^{\infty} \frac{S}{\sqrt{2\pi\sigma_N^2}} \exp\left[-\frac{(X-S)^2}{2\sigma_N^2} \right] \frac{S^{a-1}}{2\,\theta^a\,\Gamma(a)} \exp\left[-\frac{S}{\theta} \right] dS \end{aligned}$$

Making the substitution $S = -S$ in the first integral we have:

$$\begin{aligned} \text{num} &= \int_0^\infty -S \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left[-\frac{(X+S)^2}{2\sigma_N^2}\right] \frac{(S)^{a-1}}{2\,\theta^a\,\Gamma(a)} \exp\left[-\frac{S}{\theta}\right] dS \\ &+ \int_0^\infty S \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left[-\frac{(X-S)^2}{2\sigma_N^2}\right] \frac{S^{a-1}}{2\,\theta^a\,\Gamma(a)} \exp\left[-\frac{S}{\theta}\right] dS \end{aligned}$$

Expanding the exponentials and taking common factors:

$$\begin{aligned} \text{num} = \frac{\exp\left[-\frac{X^2}{2\sigma_N^2}\right]}{\sqrt{2\pi\sigma_N^2}\,2\theta^a\,\Gamma(a)} \quad \cdot \quad \Bigg[&- \int_0^\infty S^a \exp\left[-\frac{S^2}{2\sigma_N^2} - S\left(\frac{X}{\sigma_N^2} + \frac{1}{\theta}\right)\right] dS \\ &+ \int_0^\infty S^a \exp\left[-\frac{S^2}{2\sigma_N^2} - S\left(-\frac{X}{\sigma_N^2} + \frac{1}{\theta}\right)\right] dS \Bigg] \end{aligned} \tag{A.25}$$

Solving the above integrals with A.17 we get:

$$\begin{aligned} \text{num} = \frac{1}{2\sqrt{2\pi\sigma_N^2}} \frac{\sigma_N^{a+1}}{\theta^a} \frac{\Gamma(a+1)}{\Gamma(a)} \exp\left[-\frac{X^2}{2\sigma_N^2}\right] \\ \left[\exp\left[\left(\frac{\zeta_1}{2}\right)^2\right] D_{-a-1}(\zeta_1) - \exp\left[\left(\frac{\zeta_2}{2}\right)^2\right] D_{-a-1}(\zeta_2)\right] \end{aligned} \tag{A.26}$$

where $\zeta_1 = \dfrac{\sigma_N}{\theta} - \dfrac{X}{\sigma_N}, \quad \zeta_2 = \dfrac{\sigma_N}{\theta} + \dfrac{X}{\sigma_N}$

if we perform the same operations on the denominator of eq. A.24 we have:

$$\begin{aligned} \text{den} = \frac{1}{2\sqrt{2\pi\sigma_N^2}} \frac{\sigma_N^{a}}{\theta^a} \exp\left[-\frac{X^2}{2\sigma_N^2}\right] \\ \left[\exp\left[\left(\frac{\zeta_1}{2}\right)^2\right] D_{-a}(\zeta_1) + \exp\left[\left(\frac{\zeta_2}{2}\right)^2\right] D_{-a}(\zeta_2)\right] \end{aligned} \tag{A.27}$$

Dividing the numerator and the denominator we get:

$$\hat{S} = a\sigma_N \, \frac{\exp\left[\dfrac{\zeta_1^2}{4}\right] D_{-a-1}(\zeta_1) - \exp\left[\dfrac{\zeta_2^2}{4}\right] D_{-a-1}(\zeta_2)}{\exp\left[\dfrac{\zeta_1^2}{4}\right] D_{-a}(\zeta_1) + \exp\left[\dfrac{\zeta_2^2}{4}\right] D_{-a}(\zeta_2)} \tag{A.28}$$

A.5 Derivation of the MP2G estimator

The estimate of this algorithm is the value of S that maximises $\ln(p(X|S)p(S))$ where $p(X|S)$ is again given by eq. 3.5 and $p(S)$ by eq. 3.14. we consecutively have:

$$\ln(p(X|S)p(S)) = \ln\left[\frac{1}{\sqrt{2\pi\sigma_N^2}}\exp\left[-\frac{(X-S)^2}{2\sigma_N^2}\right]\frac{|S|^{a-1}}{2\theta^a\,\Gamma(a)}\exp\left[-\frac{|S|}{\theta}\right]\right] \tag{A.29}$$

Taking the derivative w.r.t. S we get:

$$\frac{d(\ln(p(X|S)p(S)))}{dS} = \frac{X-S}{\sigma_N^2} + \frac{a-1}{S} - \frac{\mathrm{sgn}(S)}{\theta} \tag{A.30}$$

Setting the above equation to zero and solving w.r.t S we get:

$$\hat{S} = \zeta + \mathrm{sgn}(X)\left[\zeta^2 + (a-1)\sigma_N^2\right]^{1/2} \qquad \text{where} \quad \zeta = \frac{X}{2} - \mathrm{sgn}(X)\frac{\sigma_N^2}{2\theta} \tag{A.31}$$

The sgn(.) in the definition of ζ comes from the fact that the maximum of the posterior density occurs at an S which has the same sign with X. The sgn(.) before the square root appears because one of the two solutions of $d(\ln(p(X|S)p(S)))/dS = 0$ is chosen according to the rules stated in appendix A.3.

A.6 Derivation of the MS1C estimator

The estimator for this algorithm can be obtained by substituting eqs. 3.20 and 3.21 into 3.22. The numerator of the last equation will then read:

$$\mathrm{num} = \int_0^\infty \int_0^{2\pi} \frac{AR}{2\pi\sigma_N^2}\exp\left[-\frac{R^2+A^2-2RA\cos(\psi-\phi)}{2\sigma_N^2}\right]\frac{2A^{a-1}\exp\left[-\frac{A^2}{\theta}\right]}{2\pi\,\theta^{a/2}\,\Gamma(a/2)}\,d\phi dA \tag{A.32}$$

which after some algebraic manipulations can be written as:

$$\mathrm{num} = K\int_0^\infty A^a \exp\left[-A^2\left(\frac{\theta+2\sigma_N^2}{\theta\,2\sigma_N^2}\right)\right] J_0\left(i\frac{RA}{\sigma_N^2}\right)\,dA \tag{A.33}$$

where

$$J_0\left(i\frac{RA}{\sigma_N^2}\right) = \frac{1}{2\pi}\int_0^{2\pi} \exp\left[\frac{RA\cos(\psi-\phi)}{\sigma_N^2}\right] d\phi \tag{A.34}$$

and $J_0(z)$ is the Bessel function of the first kind and zeroth order (see [42] eqs. 8.406.3, 8.431.5). K is:

$$K = \frac{2R}{2\pi\,\sigma_N^2\,\theta^{a/2}\,\Gamma(a/2)} \exp\left[-\frac{R^2}{2\sigma_N^2}\right] \tag{A.35}$$

The integral in eq. A.33 can be solved with formula 6.631.1 from [42] which is stated below.

$$\int_0^\infty x^\mu \mathrm{e}^{-\delta x^2} J_\nu(\beta x)\, dx = \frac{\beta^\nu\,\Gamma\left(\frac{\nu+\mu+1}{2}\right)}{2^{v+1}\,\delta^{(\mu+\nu+1)/2}\,\Gamma(\nu+1)}\ {}_1F_1\left(\frac{\nu+\mu+1}{2}, \nu+1, -\frac{\beta^2}{4\delta}\right) \tag{A.36}$$

Solving the integral we get:

$$\text{num} = K\left(\frac{2\sigma_N^2\theta}{\theta+2\sigma_N^2}\right)^{\frac{a+1}{2}}\ \frac{\Gamma(\frac{a+1}{2})}{2}\ {}_1F_1\left(\frac{a+1}{2}, 1, \frac{R^2\theta}{2\sigma_N^2(\theta+2\sigma_N^2)}\right) \tag{A.37}$$

Performing the same operations on the denominator we get:

$$\text{den} = K\left(\frac{2\sigma_N^2\theta}{\theta+2\sigma_N^2}\right)^{a/2}\ \frac{\Gamma(a/2)}{2}\ {}_1F_1\left(a/2, 1, \frac{R^2\theta}{2\sigma_N^2(\theta+2\sigma_N^2)}\right) \tag{A.38}$$

Dividing the the numerator (num) with the denominator (den) we get:

$$\hat{A} = \sqrt{2\sigma_N^2\,\zeta}\ \frac{\Gamma(\frac{a+1}{2})}{\Gamma(a/2)}\ \frac{{}_1F_1(\frac{a+1}{2}, 1, \frac{R^2}{2\sigma_N^2}\zeta)}{{}_1F_1(a/2, 1, \frac{R^2}{2\sigma_N^2}\zeta)} \qquad \text{where} \quad \zeta = \frac{\theta}{\theta+2\sigma_N^2} \tag{A.39}$$

A.7 Derivation of the MP1C estimator

Substituting $p(R,\psi|A,\phi)$ and $p(A)$ from 3.20 and 3.21 and $p(\phi)=\frac{1}{2\pi}$ into eq. 3.25 yields:

$$\hat{A} = \arg\max_{A} \ln\left[\int_0^{2\pi} \frac{R}{2\pi\sigma_N^2}\exp\left[-\frac{R^2+A^2-2RA\cos(\psi-\phi)}{2\sigma_N^2}\right] \cdot \frac{2A^{a-1}}{2\pi\,\theta^{a/2}\,\Gamma(a/2)}\exp\left[-\frac{A^2}{\theta}\right] d\phi\right] \quad \text{(A.40)}$$

After some rearrangement the logarithm can be written as:

$$\ln\left[\frac{2R}{2\pi\sigma_N^2\,\theta^{a/2}\,\Gamma(a/2)}A^{a-1}\exp\left[-\frac{R^2+A^2}{2\sigma_N^2}-\frac{A^2}{\theta}\right]\frac{1}{2\pi}\int_0^{2\pi}\exp\left[\frac{RA\cos(\psi-\phi)}{\sigma_N^2}\right]d\phi\right] \quad \text{(A.41)}$$

Using eq. 8.431.5 from [42] we have:

$$I_0\left(\frac{RA}{\sigma_N^2}\right) = \frac{1}{2\pi}\int_0^{2\pi}\exp\left[\frac{RA\cos(\psi-\phi)}{\sigma_N^2}\right]d\phi \quad \text{(A.42)}$$

where $I_0(z)$ is the modified Bessel function of the first kind. Using also the approximation [73]

$$I_0(z) \sim \mathrm{e}^z/\sqrt{2\pi z} \quad \text{(A.43)}$$

the logarithm in eq. A.41 can be written as:

$$\ln\left[\frac{2R}{2\pi\sigma_N^2\,\theta^{a/2}\,\Gamma(a/2)}A^{a-1}\exp\left[-\frac{R^2+A^2}{2\sigma_N^2}-\frac{A^2}{\theta}\right]\frac{\exp\left[\frac{RA}{\sigma_N^2}\right]}{\sqrt{2\pi\frac{RA}{\sigma_N^2}}}\right] \quad \text{(A.44)}$$

Taking the derivative of the above expression w.r.t. A, setting to zero and solving w.r.t A we get:

$$\hat{A} = \zeta\frac{R}{2} \pm \left[\left(\zeta\frac{R}{2}\right)^2 + (a-1.5)\,\sigma_N^2\,\zeta\right]^{1/2} \quad \text{where} \quad \zeta = \frac{\theta}{\theta+2\sigma_N^2} \quad \text{(A.45)}$$

From the above two solutions the valid is the one which is a maximum and positive.

Some further analysis shows that this is always the one with the (+).

A.8 Derivation of the MS1G estimator

The estimator for this algorithm can be obtained by substituting eqs. 3.20 and 3.28 into 3.22. The numerator of the last equation will then read:

$$\text{num} = \int_0^{\infty} \int_0^{2\pi} \frac{AR}{2\pi\sigma_N^2} \exp\left[-\frac{R^2 + A^2 - 2RA\cos(\psi - \phi)}{2\sigma_N^2}\right] \frac{A^{a-1} \exp\left[-\frac{A}{\theta}\right]}{2\pi\, \theta^a\, \Gamma(a)}\, d\phi dA \tag{A.46}$$

which after some algebraic manipulations can be written as:

$$\text{num} = K \int_0^{\infty} A^a \exp\left[-\frac{A^2}{2\sigma_N^2} - \frac{A}{\theta}\right] I_0\left(\frac{RA}{\sigma_N^2}\right) dA \tag{A.47}$$

where $I_0\left(\frac{RA}{\sigma_N^2}\right)$ is defined in eq. A.42 and K is given by:

$$K = \frac{R}{2\pi\, \sigma_N^2\, \theta^a\, \Gamma(a)} \exp\left[-\frac{R^2}{2\sigma_N^2}\right] \tag{A.48}$$

Performing the same operations on the denominator we get:

$$\text{den} = K \int_0^{\infty} A^{a-1} \exp\left[-\frac{A^2}{2\sigma_N^2} - \frac{A}{\theta}\right] I_0\left(\frac{RA}{\sigma_N^2}\right) dA \tag{A.49}$$

where K is given by eq. A.48. Division of the eqs. A.47 and A.49 yields the expression given in eq. 3.29

A.9 Derivation of the MP1G estimator

Substituting $p(R,\psi|A,\phi)$ and $p(A)$ from 3.20 and 3.28 and $p(\phi)=\frac{1}{2\pi}$ into eq. 3.25 yields:

$$\hat{A} = \arg\max_{A} \ln\left[\int_0^{2\pi} \frac{R}{2\pi\sigma_N^2}\exp\left[-\frac{R^2+A^2-2RA\cos(\psi-\phi)}{2\sigma_N^2}\right] \cdot \frac{A^{a-1}}{2\pi\,\theta^a\,\Gamma(a)}\exp\left[-\frac{A}{\theta}\right]\,d\phi\right] \quad \text{(A.50)}$$

After some rearrangement the logarithm can be written as:

$$\ln\left[\frac{R}{2\pi\sigma_N^2\,\theta^a\,\Gamma(a)}A^{a-1}\exp\left[-\frac{R^2+A^2}{2\sigma_N^2}-\frac{A}{\theta}\right]\frac{1}{2\pi}\int_0^{2\pi}\exp\left[\frac{RA\cos(\psi-\phi)}{\sigma_N^2}\right]\,d\phi\right] \quad \text{(A.51)}$$

Transforming the integral as in appendix A.7 the above expression becomes:

$$\ln\left[\frac{R}{2\pi\sigma_N^2\,\theta^a\,\Gamma(a)}A^{a-1}\exp\left[-\frac{R^2+A^2}{2\sigma_N^2}-\frac{A}{\theta}\right]\frac{\exp\left[\frac{RA}{\sigma_N^2}\right]}{\sqrt{2\pi\frac{RA}{\sigma_N^2}}}\right] \quad \text{(A.52)}$$

Taking the derivative of the above expression w.r.t. A, setting to zero and solving w.r.t A we get:

$$\hat{A}=\zeta\pm\left[\zeta^2+(a-1.5)\sigma_N^2\right]^{1/2} \text{ where } \quad \zeta=\frac{R}{2}-\frac{\sigma_N^2}{2\theta} \quad \text{(A.53)}$$

From the above two solutions the valid is the one with the (+) because it is always positive and a maximum.

A.10 Derivation of the MP1L estimator

Substituting eqs. 3.20, 3.33 and $p(\phi) = \frac{1}{2\pi}$ into eq. 3.25 yields:

$$\hat{A} = \arg\max_{A} \ln \left[\int_0^{2\pi} \frac{R}{2\pi\sigma_N^2} \exp\left[- \frac{R^2 + A^2 - 2RA\cos(\psi - \phi)}{2\sigma_N^2} \right] \cdot \frac{1}{2\pi} \frac{\sqrt{a}}{\sqrt{\pi}\, A} \exp\left[-a\ (\ln(A) - \theta)^2 \right] d\phi \right] \quad \text{(A.54)}$$

Discarding the terms that are constant with respect to A and rearranging the remaining ones we have:

$$\hat{A} = \arg\max_{A} \ln \left[\frac{1}{A} \exp\left[-\frac{R^2 + A^2}{2\sigma_N^2} \right] \exp\left[-a\ (\ln(A) - \theta)^2 \right] \cdot \frac{1}{2\pi} \int_0^{2\pi} \exp\left[\frac{2RA\cos(\psi - \phi)}{2\sigma_N^2} \right] \right] \quad \text{(A.55)}$$

Using eq. A.42 the above equation can be written as:

$$\hat{A} = \arg\max_{A} \left[-\ln(A) - \frac{R^2 + A^2}{2\sigma_N^2} - a\ (\ln(A) - \theta)^2 + \ln\left(I_0\left(\frac{2RA}{2\sigma_N^2} \right) \right) \right] \quad \text{(A.56)}$$

Appendix B

Amplitude density functions and their logarithmic transformation

	$p(A)$	$p(y), \quad y = \ln(A)$
Chi	$\frac{2}{\theta^{a/2}\Gamma(a/2)} \; A^{a-1} \; \exp\left[-\frac{A^2}{\theta}\right]$	$\frac{2}{\theta^{a/2}\Gamma(a/2)} \; \exp\left[-\frac{\mathrm{e}^{2y}}{\theta} + ya\right]$
Gamma	$\frac{1}{\theta^{a}\Gamma(a)} \; A^{a-1} \; \exp\left[-\frac{A}{\theta}\right]$	$\frac{1}{\theta^{a}\Gamma(a)} \; \exp\left[-\frac{\mathrm{e}^{y}}{\theta} + ya\right]$
Lognormal	$\frac{\sqrt{a}}{\sqrt{\pi}A} \; \exp\left[-a(\ln(A) - \theta)^2\right]$	$\frac{\sqrt{a}}{\sqrt{\pi}} \; \exp\left[-a(y - \theta)^2\right]$

Table B.1: Amplitude density functions and their logarithmic transformation.

Appendix C

The effect of using long term priors to speech quality

In this appendix we investigate the effect of using the values of the priors' parameters that have been estimated using long term speech data. In particular, we demonstrate the effect of using the values that have been estimated using the all the available speech STFT data, as discussed in §4.1, and make a comparison with the case when the same values of a are used, but the scale parameter θ is estimated using the DD method.

The use of the scale parameter value estimated from the long term priors compromises the suppression of the residual noise and, perhaps more importantly, results in a residual noise that has a strong musical character. The incorporation of the DD method on the other hand, suppresses the residual noise more effectively, and smooths the spurious spectral peaks, thus making the residual noise more uniform. A downside of the DD method is that some of the speech spectral components are also suppressed. This drawback however is outweighted by the lower level of the residual noise and its more uniform character. The spectrograms of figure C.1 illustrate the above observations. The utterance described in §5.2 is enhanced with the MP1G and MS1G algorithms with either fixed or adaptive values of θ. The value of a in both cases is 0.28, as it was estimated in §4.1.

Table C.1 shows the results in the objective measures for all the MMSE algorithms with either fixed or adaptively estimated values of θ via the DD method. Table C.2

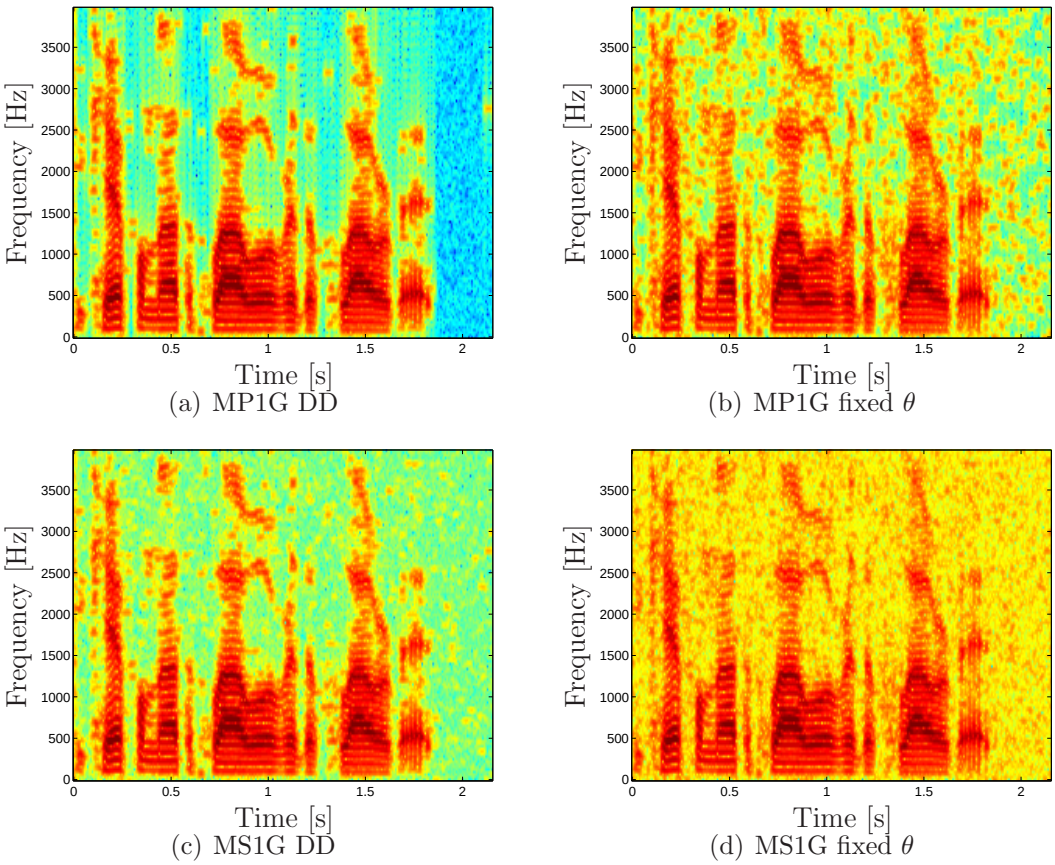


Figure C.1: Speech enhanced with the MP1G and MS1G algorithms. The scale parameter θ was estimated either with the DD method, or the fixed values estimated in §4.1 were used.

shows the respective results for the MAP algorithms. For the MMSE algorithms the objective measures favour the DD method for the majority of the cases, especially in the low input SegSNR conditions, where the effect of the background noise is more damaging to the quality of speech. For the MAP algorithms however, it is interesting to note that although the SegSNR favours the DD method for the majority of the cases, the best scores for the PESQ are achieved with the fixed values of θ. This was rather striking, since both informal listening tests and the examination of spectrograms indicated that the DD method resulted in lower levels of residual noise, which also has a significantly more uniform character, and there was no apparent aspect of speech quality in which the results obtained with the fixed values of θ surpassed those obtained using the DD method.

A possible explanation for this observation could be the following: for the values of a estimated using the long term priors, which are used in this appendix, the MAP

	White Noise				Car Noise			
	SegSNR		PESQ		SegSNR		PESQ	
	DD	Fixed	DD	Fixed	DD	Fixed	DD	Fixed
	0 dB Input SegSNR							
MS1C	7.76	5.68	2.88	2.65	10.93	8.43	3.54	3.34
MS2C	7.33	5.41	2.83	2.62	10.53	8.57	3.48	3.33
MS1G	7.79	5.31	2.87	2.58	10.95	9.00	3.54	3.30
MS2G	7.45	5.48	2.82	2.60	10.65	9.68	3.48	3.33
MS1L	7.77	5.30	2.82	2.57	11.05	9.08	3.47	3.27
	10 dB Input SegSNR							
MS1C	13.82	13.51	3.51	3.32	17.28	15.75	3.97	3.92
MS2C	13.20	13.30	3.45	3.29	16.75	15.58	3.93	3.90
MS1G	13.91	13.12	3.51	3.22	17.34	15.59	3.97	3.85
MS2G	13.40	13.01	3.44	3.21	16.91	15.91	3.93	3.85
MS1L	13.65	12.61	3.38	3.10	17.14	15.59	3.93	3.73
	20 dB Input SegSNR							
MS1C	21.22	21.63	4.00	3.95	24.50	23.89	4.26	4.27
MS2C	20.46	21.46	3.94	3.92	23.79	23.71	4.24	4.26
MS1G	21.35	21.50	3.99	3.84	24.58	23.49	4.26	4.25
MS2G	20.73	21.38	3.94	3.82	24.01	23.49	4.24	4.24
MS1L	21.05	21.02	3.92	3.62	24.29	23.08	4.25	4.17

Table C.1: Objective measures' scores for the MMSE algorithms, using adaptive or fixed values of θ.

algorithms are not well defined. This implies that for a large number of samples, especially in the noise dominated areas of the spectrograms, estimates do not exist, and the strategy we follow is to simply suppress the noisy samples by a fixed amount [1]. Consequently the difference between the residual noise levels obtained with the fixed and the adaptive values of θ is smaller for the MAP compared to the MMSE algorithms. This characteristic in combination with the fact that some weak speech spectral components are better preserved with the fixed values of θ might give rise to the better scores of the PESQ measure for the MAP algorithms. Nevertheless, we believe that the above behaviour of the PESQ measure is an indication of its weakness in penalising the musical character of the residual noise.

[1] see the derivation of the MAP algorithms in chapter 3

	White Noise				Car Noise			
	SegSNR		PESQ		SegSNR		PESQ	
	DD	Fixed	DD	Fixed	DD	Fixed	DD	Fixed
	0 dB Input SegSNR							
MP1C	7.67	5.81	2.70	2.66	11.11	9.50	3.42	3.38
MP2C	7.21	5.15	2.73	2.61	10.50	8.91	3.40	3.34
MP1G	7.69	6.91	2.64	2.71	11.11	10.78	3.40	3.46
MP2G	7.26	6.10	2.67	2.69	10.54	10.26	3.36	3.41
MP1L	6.84	7.52	2.36	2.72	10.04	10.80	3.19	3.39
	10 dB Input SegSNR							
MP1C	13.56	13.39	3.32	3.34	17.13	16.23	3.91	3.93
MP2C	13.07	13.01	3.35	3.28	16.64	15.63	3.90	3.89
MP1G	13.55	13.52	3.27	3.36	17.10	16.96	3.90	3.94
MP2G	13.04	13.08	3.27	3.29	16.62	16.47	3.87	3.90
MP1L	12.48	13.77	3.01	3.44	16.21	17.04	3.77	3.95
	20 dB Input SegSNR							
MP1C	20.97	21.49	3.92	3.95	24.34	24.07	4.24	4.25
MP2C	20.36	21.26	3.91	3.90	23.68	23.65	4.24	4.24
MP1G	20.98	21.49	3.90	3.95	24.31	24.28	4.24	4.25
MP2G	20.35	21.25	3.86	3.88	23.65	23.83	4.23	4.24
MP1L	19.77	21.53	3.69	3.95	23.22	24.37	4.18	4.25

Table C.2: Objective measures' scores for the MAP algorithms, using adaptive or fixed values of θ.

Appendix D

Derivation of the the joint Chi MRF density

The conditional density of the Chi MRF is

$$p(x_i|x_{\mathrm{n}(i)}) \propto x_i^{a-1} \exp\left[-\frac{1}{\theta_i}\left(x_i - \sum_{j\in \mathrm{n}(i)} b_{ij}x_j\right)^2\right] \tag{D.1}$$

The logarithm of the factorisation in eq. 7.6 can be written as

$$\begin{aligned}\ln\left(\frac{p(x)}{p(z)}\right) &= \sum_{i\in Q}\left[\ln\left(p(x_i|x_1,...,x_{i-1},z_{i+1},...,z_q)\right)\right] \\ &- \sum_{i\in Q}\left[\ln\left(p(z_i|x_1,...,x_{i-1},z_{i+1},...,z_q)\right)\right]\end{aligned} \tag{D.2}$$

Substituting eq. D.1 in the above factorisation we have

$$\begin{aligned}\ln\left(\frac{p(x)}{p(z)}\right) &= \sum_{i\in Q}\left[(a-1)\ln(x_i) - \frac{x_i^2}{\theta_i} + 2\frac{x_i\Omega}{\theta_i} + \frac{\Omega^2}{\theta_i}\right] \\ &- \sum_{i\in Q}\left[(a-1)\ln(z_i) + \frac{z_i^2}{\theta_i} - 2\frac{z_i\Omega}{\theta_i} - \frac{\Omega^2}{\theta_i}\right]\end{aligned} \tag{D.3}$$

where

$$\Omega = \sum_{\{j\in \mathrm{n}(i):j<i\}} b_{ij}\, x_j + \sum_{\{j\in \mathrm{n}(i):j>i\}} b_{ij}\, z_j$$

Elimination of the $\frac{\Omega^2}{\theta_i}$ terms in eq. D.3 yields

$$\begin{aligned}\ln\left(\frac{p(x)}{p(z)}\right) &= \sum_{i\in Q}\left[\ln(x_i^{a-1}) - \frac{x_i^2}{\theta_i} + \sum_{\{j\in \mathrm{n}(i):j<i\}} \frac{2b_{ij}}{\theta_i}\, x_i x_j + \sum_{\{j\in \mathrm{n}(i):j>i\}} \frac{2b_{ij}}{\theta_i}\, x_i z_j\right] \\ &- \sum_{i\in Q}\left[\ln(z_i^{a-1}) + \frac{z_i^2}{\theta_i} - \sum_{\{j\in \mathrm{n}(i):j<i\}} \frac{2b_{ij}}{\theta_i}\, z_i x_j - \sum_{\{j\in \mathrm{n}(i):j>i\}} \frac{2b_{ij}}{\theta_i}\, z_i z_j\right]\end{aligned}$$

Under the assumption that $\frac{b_{ij}}{\theta_i} = \frac{b_{ji}}{\theta_j}$ and $j \in \mathrm{n}(i)$ if and only if $i \in \mathrm{n}(j)$ it holds that

$$\sum_{i\in Q}\sum_{\{j\in \mathrm{n}(i):j>i\}} \frac{2b_{ij}}{\theta_i}\, x_i z_j = \sum_{i\in Q}\sum_{\{j\in \mathrm{n}(i):j<i\}} \frac{2b_{ij}}{\theta_i}\, z_i x_j$$

The expression for $\ln(p(x))$ then becomes

$$\begin{aligned}\ln\left(p(x)\right) &\propto \sum_{i\in Q}\left[\ln(x_i^{a-1}) - \frac{x_i^2}{\theta_i} + \sum_{\{j\in \mathrm{n}(i):j<i\}} \frac{2b_{ij}}{\theta_i}\, x_i x_j\right] \\ p(x) &\propto \prod_{i\in Q}\left(x_i^{a-1}\right)\exp\left[\sum_{i\in Q}\left[-\frac{x_i^2}{\theta_i} + \sum_{\{j\in \mathrm{n}(i):j<i\}} \frac{2b_{ij}}{\theta_i}\, x_i x_j\right]\right]\end{aligned}$$

Noting also that

$$\sum_{i\in Q}\sum_{\{j\in \mathrm{n}(i):j<i\}} \frac{2b_{ij}}{\theta_i}\, x_i x_j = \sum_{\{i,j\}\in C} \frac{2b_{ij}}{\theta_i}\, x_i x_j$$

where C is the unordered set of pairs of indices i, j such that $\{i, j\} \in C$ if and only if x_i and x_j are neighbours. The joint density $p(x)$ can be finally written as

$$p(x) \propto \prod_{i\in Q}\left(x_i^{a-1}\right)\exp\left[-\sum_{i\in Q}\frac{x_i^2}{\theta_i} + \sum_{\{i,j\}\in C} \frac{2b_{ij}}{\theta_i}\, x_i x_j\right] \tag{D.4}$$

In order for the above expression to constitute a valid probability density function we also require that $|\int_x p(x)dx| < \infty$. This condition is satisfied if the argument of the exponential is negative for all the possible values of x. The values of the

parameters b_{ij} and θ_i that satisfy this condition can be found if we write eq. D.4 as

$$p(x) \propto \prod_{i \in Q} \left(x_i^{a-1}\right) \exp\left[-\sum_{i \in Q} \frac{\left(1 - \sum_{j \in \mathrm{n}(i)} b_{ij}\right)}{\theta_i} x_i^2 - \sum_{\{i,j\} \in C} \frac{b_{ij}}{\theta_i} \left(x_i - x_j\right)^2\right] \tag{D.5}$$

The above equation reveals that the argument of the exponential is negative for all x if $b_{ij} > 0,\ \forall i, j \in Q$ and if $\sum_{j \in \mathrm{n}(i)} b_{ij} < 1,\ \forall i \in Q$.

An alternative method for proving the validity of the joint Chi MRF density function is given by Gershgorin's circle theorem [41]. We first write eq. D.4 as

$$p(x) \propto \prod_{i \in Q} \left(x_i^{a-1}\right) \exp\left[-xGx^{\mathrm{T}}\right] \tag{D.6}$$

where the elements of the matrix G are $G_{ii} = 1/\theta_i$ and $G_{ij} = -b_{ij}/\theta_i$. The MRF defined by $p(x)$ is valid if the matrix G is positive definite, or equivalently, if all of its eigenvalues are positive. Gershgorin's circle theorem says that the eigenvalues of a matrix G lie in circles, which are centered at the points G_{ii} on the complex plane, and whose radius is less or equal to $\sum_{j \in Q} |G_{ij}|$. If $b_{ij} > 0,\ \forall i, j \in Q$ and if $\sum_{j \in \mathrm{n}(i)} b_{ij} < 1,\ \forall i \in Q$, then Gershgorin's circles lie on the right hand side of the complex plane. Therefore, the matrix G positive definite.

References

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions.* New York: Dover, 1965.

[2] B. I. Andia, "Restoration of speech signals contaminated by stationary tones using an image perspective," *in Proc. 31st IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-06*, vol. 3, pp. 61–64, 2006.

[3] I. Andrianakis, "Speaker identification using Independent Component Analysis," *MSc Thesis, ISVR, University of Southampton*, 2004.

[4] I. Andrianakis and P. R. White, "Bayesian algorithms for speech enhancement," *ISVR Technical Report, No 305*, 2006.

[5] ——, "Noise estimation based on matching the moments of the Rayleigh distribution for speech enhancement," *in Proc. Hellenic Inst. of Acoustics 2006*, pp. 91–96, 2006.

[6] ——, "On the application of Markov Random Fields to speech enhancement," *in Proc. IMA Int. Conf. Mathematics in Signal processing*, pp. 198–201, 2006.

[7] ——, "MMSE speech spectral amplitude estimators with Chi and Gamma speech priors," *in Proc. 31st IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-06*, vol. 3, pp. 1068–1071, 2006.

[8] B. Barrowes, "Matlab routines for computation of special functions," *http://ceta.mit.edu/comp_spec_func/*, 2006.

[9] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *in Proc. 4th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-79*, vol. 4, pp. 208–211, 1979.

[10] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society, Series B*, no. 36, pp. 192–236, 1974.

[11] ——, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society, Series B*, no. 48, pp. 259–302, 1986.

[12] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian Mixture and Hidden Markov Models," *Tech. Rep. TR-97-021, ICSI*, 1998.

[13] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[14] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. Image Processing*, vol. 2, no. 3, pp. 296–310, 1993.

[15] D. R. Brillinger, "Fourier analysis of stationary processes," *Proceedings of the IEEE*, vol. 62, no. 12, pp. 1628–1643, 1974.

[16] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[17] P. Clifford, "Markov random fields in statistics," *Disorder in Physical Systems. A Volume in honour of John M. Hammersley*, pp. 19–32, 1990.

[18] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, 2002.

[19] ——, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[20] ——, "On the decision-directed estimation approach of Ephraim and Malah," *in Proc. 29th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-04*, vol. 1, pp. I–293–6, 2004.

[21] ——, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, 2005.

[22] ——, "Speech enhancement using supergaussian speech models and noncausal a priori SNR estimation," *Speech Communication*, vol. 47, no. 3, pp. 336–350, 2005.

[23] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, 2002.

[24] T. H. Dat, K. Takeda, and F. Itakura, "Generalized Gamma modeling of speech and its online estimation for speech enhancement," *in Proc. 30th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-05*, vol. 4, pp. 181–184, 2005.

[25] J. R. J. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York, NY: Macmillan Publishing Company, 1993.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[27] M. Dendrinos, S. Bakamidis, and G. Carayiannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–57, 1991.

[28] G. H. Ding, T. Huang, and B. Xu, "Suppression of additive noise using a power spectral density MMSE estimator," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 585–588, 2004.

[29] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," *EUROSPEECH '95*, pp. 1513–1516, 1995.

[30] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[31] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[32] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[33] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[34] J. M. Fadili and L. Boubchir, "Analytical form for a Bayesian wavelet estimator of images using the Bessel k form densities," *IEEE Trans. Image Processing*, vol. 14, no. 2, pp. 231–240, 2005.

[35] M. A. T. Figueiredo and J. M. N. Leitao, "Unsupervised image restoration and edge location using compound Gauss-Markov random fields and the MDL principle," *IEEE Trans. Image Processing*, vol. 6, no. 8, pp. 1089–1102, 1997.

[36] W. Gander and W. Gautschi, "Adaptive quadrature - revisited," *BIT*, vol. 40, no. 1, pp. 84–101, 2000.

[37] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, 1998.

[38] S. Gazor and W. Zhang, "Speech enhancement employing Laplacian-Gaussian mixture," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 896–904, 2005.

[39] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, pp. 721–741, 1984.

[40] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Processing*, vol. 39, no. 8, pp. 1732–1742, 1991.

[41] G. H. Golub and C. F. Van Loan, *Matrix computations*. Baltimore: Johns Hopkins University Press, third edition, 1996.

[42] I. S. Gradshteyn and I. W. Ryzhik, *Tables of Integrals, Series and Products*. New York: Academic Press, 1965.

[43] G. Gravier, M. Sigelle, and G. Chollet, "A Markov random field based multi-band model," *in Proc. 25th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-00*, vol. 3, pp. 1619–1622, 2000.

[44] J. M. Hammersley and P. Clifford, "Markov random fields on finite graphs and lattices," *(Unpublished)*, 1971.

[45] J. H. L. Hansen, V. Radhakrishnan, and K. H. Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2049–2063, 2006.

[46] R. C. Hendriks, R. Heusdens, and J. Jensen, "Adaptive time segmentation for improved speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2064–2074, 2006.

[47] ——, "Speech enhancement under a combined stochastic-deterministic model," *in Proc. 31st IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-06*, vol. 1, pp. I–453, 2006.

[48] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," *in Proc. 20th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-95*, vol. 1, pp. 153–156, 1995.

[49] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," *in Proc. 27th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-02*, vol. 1, pp. I–573, 2002.

[50] ——, "A perceptually motivated approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 457–465, 2003.

[51] ——, "Evaluation of objective measures for speech enhancement," *INTERSPEECH 2006*, 2006.

[52] ——, "Subjective comparison of speech enhancement algorithms," *in Proc. 31st IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-06*, vol. 1, pp. I–153, 2006.

[53] *Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, ITU-T Recommendation P.862 Std., Feb. 2001.

[54] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 700–708, 2003.

[55] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 6, pp. 439–448, 1995.

[56] N. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*. New York: John Wiley & Sons, 1994, vol. 1.

[57] A. C. Kokaram and S. J. Godsill, "A system for reconstruction of missing data in image sequences using sampled 3D AR models and MRF motion priors," *Europ. Conf. Computer and Vision (ECCV)*, pp. 613–624, 1996.

[58] ——, "MCMC for joint noise reduction and missing data treatment in degraded video," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 189–205, 2002.

[59] A. Kraskov, H. Stogbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, no. 066138, 2004.

[60] S. Kullback, *Information Theory and Statistics*. New York: John Wiley and Sons, 1959.

[61] S. M. Kuo and J. Kunduru, "Subband adaptive noise canceler for hands-free cellular phone applications," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*, pp. 19–22, 1993.

[62] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Processing Letters*, vol. 10, no. 4, pp. 104–106, 2003.

[63] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[64] L. Lin, W. H. Holmes, and E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement," *Electronics Letters*, vol. 39, no. 9, pp. 754–755, 2003.

[65] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005.

[66] T. Lotter and P. Vary, "Noise reduction by maximum a posteriori spectral amplitude estimation with supergaussian speech modeling," *International Workshop on Acoustic Echo and Noise Control (IWAENC2003),*, pp. 83–86, 2003.

[67] ——, "Noise reduction by joint maximum a posteriori spectral amplitude and phase estimation with super-gaussian speech modelling," *in Proc. of EUSIPCO-04 (Vienna, Austria),*, pp. 1457–1460, 2004.

[68] N. Ma, M. Bouchard, and R. A. Goubran, "Speech enhancement using a masking threshold constrained Kalman filter and its heuristic implementations," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 19–32, 2006.

[69] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," *in Proc. 24th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-99*, vol. 2, pp. 789–792, 1999.

[70] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. Eur. Signal Processing Conf.(EUSIPCO '94)*, pp. 1182–1185, 1994.

[71] ——, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing,*, vol. 9, pp. 504–512, 2001.

[72] ——, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, 2005.

[73] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[74] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 2, pp. 159–167, 2000.

[75] D. S. Moore and G. P. McCabe, *Introduction to the practice of statistics.* New York: W.H.Freeman & Co., 1989.

[76] S. Nordebo, S. Nordholm, B. Bengtsson, and I. Claesson, "Noise reduction using an adaptive microphone array in a car - a speech recognition evaluation," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 16–18, 1993.

[77] T. P. O'Rourke and R. L. Stevenson, "Improved image decompression for reduced transform coding artifacts," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 5, no. 6, pp. 490–499, 1995.

[78] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *in Proc. 12th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-87*, vol. 12, pp. 177–180, 1987.

[79] A. Papoulis, *Probability, random variables and stochastic processes.* New York: McGraw Hill, 1965.

[80] P. Pèrez, "Markov random fields and images," *CWI Quarterly*, vol. 11, no. 4, pp. 413–437, 1998.

[81] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," *in Proc. 9th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-84*, vol. 9, pp. 53–56, 1984.

[82] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: the art of scientific computing.* New York: Cambridge University Press, 1992.

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

[84] S. Rangachari, P. C. Loizou, and Y. Hu, "A noise estimation algorithm with rapid adaptation for highly nonstationary environments," *in Proc. 29th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-04*, vol. 1, pp. I–8, 2004.

[85] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 2, pp. 87–95, 2001.

[86] C. Ris and S. Dupont, "Assessing local noise level estimation methods: Application to noise robust ASR," *Speech Communication*, vol. 34, pp. 141–158, 2001.

[87] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," *in Proc. 26th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-01*, vol. 2, pp. 749–752, 2001.

[88] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *in Proc. 21st IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-96*, vol. 2, pp. 629–632, 1996.

[89] J. W. Shin, J. H. Chang, and N. S. Kim, "Statistical modeling of speech signals based on generalized Gamma distribution," *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 258–261, 2005.

[90] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 4, pp. 328–337, 1998.

[91] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[92] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," *in Proc. 25th IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-00*, vol. 3, pp. 1875–1878, 2000.

[93] L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree, "MELP: the new Federal standard at 2400 bps," *in Proc. 22nd IEEE Int. Conf. Acoust. Speech Signal Process., ICASSP-97*, vol. 2, pp. 1591–1594, 1997.

[94] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 6, pp. 497–514, 1997.

[95] D. Van Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Computer Speech and Language*, vol. 3, no. 2, pp. 151–168, 1989.

[96] H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I.* New York: John Wiley & Sons, 1968.

[97] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, 1999.

[98] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[99] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim Malah suppression rule for audio signal enhancement," *EURASIP Journal on Applied Signal Processing*, vol. 10, pp. 1043–1051, 2003.

[100] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 4, pp. 475–486, 2005.

[101] ——, "Masking β-order MMSE speech enhancement," *Speech Communication*, vol. 48, pp. 57–70, 2005.

[102] E. Zavarehei, S. Vaseghi, and Q. Yan, "Inter-frame modelling of DFT trajectories of speech and noise for speech enhancement using Kalman filters," *Speech Communication*, vol. 48, no. 11, pp. 1545–1555, 2006.