Bayesian short-time spectral amplitude estimators for single-channel speech enhancement

Eric Plourde

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

October 2009

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

Single-channel speech enhancement algorithms are used to remove background noise in speech. They are present in many common devices such as cell phones and hearing aids. In the Bayesian short-time spectral amplitude (STSA) approach for speech enhancement, an estimate of the clean speech STSA is derived by minimizing the statistical expectation of a chosen cost function. Examples of such estimators are the minimum mean square error (MMSE) STSA, the β-order MMSE STSA (β-SA), which includes a power law parameter, and the weighted Euclidian (WE), which includes a weighting parameter.

This thesis analyzes single-channel Bayesian STSA estimators for speech enhancement with the aim of, firstly, gaining a better understanding of their properties and, secondly, proposing new cost functions and statistical models to improve their performance. In addition to a novel analysis of the β-SA estimator for parameter $\beta \leq 0$, three new families of estimators are developed in this thesis: the Weighted β-SA (Wβ-SA), the Generalized Weighted family of STSA estimators (GWSA) and a family of multi-dimensional Bayesian STSA estimators.

The Wβ-SA combines the power law of the β-SA and the weighting factor of the WE. Its parameters are chosen based on the characteristics of the human auditory system which is found to have the advantage of improving the noise reduction at high frequencies while limiting the speech distortions at low frequencies. An analytical generalization of a cost function structure found in many existing Bayesian STSA estimators is proposed through the GWSA family of estimators. This allows a unification of Bayesian STSA estimators and, moreover, provides a better understanding of this general class of estimators. Finally, we propose a multi-dimensional family of estimators that accounts for the correlated frequency components in a digitized speech signal. In fact, the spectral components of the clean speech are traditionally assumed uncorrelated in Bayesian STSA estimators, however, this assumption is inexact since some correlation is present in practice. Objective and subjective experiments are performed in different noise environments and at several signal-to-noise ratios (SNR). Results show the superiority of the proposed estimators over benchmark estimators.

Sommaire

Les algorithmes de rehaussement de la parole à voie unique sont utilisés afin de réduire le bruit de fond d'un signal de parole bruité. Ils sont présents dans plusieurs appareils tels que les téléphones sans fil et les prothèses auditives. Dans l'approche bayésienne d'estimation de l'amplitude spectrale locale (*Short-Time Spectral Amplitude* - STSA) pour le rehaussement de la parole, un estimé de la STSA non bruitée est déterminé en minimisant l'espérance statistique d'une fonction de coût. Ce type d'estimateurs incluent le MMSE STSA, le β-SA, qui intègre un exposant comme paramètre de la fonction de coût, et le WE, qui possède un paramètre de pondération.

Cette thèse étudie les estimateurs bayésiens du STSA avec pour objectifs d'approfondir la compréhension de leurs propriétés et de proposer de nouvelles fonctions de coût ainsi que de nouveaux modèles statistiques afin d'améliorer leurs performances. En plus d'une étude approfondie de l'estimateur β-SA pour les valeurs de $\beta \leq 0$, trois nouvelles familles d'estimateur sont dévelopées dans cette thèse: le β-SA pondéré (*Weighted β-SA* - Wβ-SA), une famille d'estimateur du STSA généralisé et pondéré (*Generalized Weighted STSA* - GWSA) ainsi qu'une famille d'estimateur du STSA multi-dimensionnel.

Le Wβ-SA combine l'exposant présent dans le β-SA et le paramètre de pondération du WE. Ses paramètres sont choisis en considérant certaines caractéristiques du système auditif humain ce qui a pour avantage d'améliorer la réduction du bruit de fond à hautes fréquences tout en limitant les distorsions de la parole à basses fréquences. Une généralisation de la structure commune des fonctions de coût de plusieurs estimateurs bayésiens du STSA est proposée à l'aide de la famille d'estimateur GWSA. Cette dernière permet une unification des estimateurs bayésiens du STSA et apporte une meilleure compréhension de cette classe générale d'estimateur. Finalement, une nouvelle famille d'estimateurs multi-dimensionnels qui permet de considérer les corrélations présentes entre les composantes fréquentielles d'un signal numérisé de parole est proposée. En effet, les composantes spectrales du signal de parole non-bruité sont traditionnellement assumées comme étant non-correlées dans l'approche d'estimation bayésienne du STSA, toutefois, cette hypothèse est inexacte puisqu'il existe dans les faits une corrélation entre les différentes composantes spectrales d'un signal numérique de parole. Des expériences de type subjectif et objectif sont effectuées pour plusieurs rapports signal-sur-bruit ainsi que différents types de bruit. Les résultats démontrent la supériorité des estimateurs proposés par rapport à ceux comparés.

Acknowledgments

First and foremost, I would like to thank Professor Benoît Champagne for his guidance, based on strong scientific rigor and a much appreciated human approach. I would also like to thank him for the many great opportunities he gave me that have all enhanced my education including giving lectures, attending international conferences or performing article reviews.

I am also grateful for the financial support provided by a scholarship from the *Fonds québécois de la recherche sur la nature et les technologies* (FQRNT), by Professor Champagne via research grants from the Natural Sciences and Engineering Research Council of Canada (CRSNG) and by the Loans and Bursaries Program of the Government of Quebec without which this thesis would not have been possible.

This journey at McGill would not have been the same without the many interesting discussions with my fellow members of the Telecommunications & Signal Processing Laboratory, in particular: Wei, Benoît, Patrick, Bo, Francois, Mohamed and Joe.

I would also like to acknowledge the role of Professor Yves Gagnon, who, many years ago gave me my first opportunity to perform academic research at the Université de Moncton and has been a strong inspiration in setting myself hard achievable professional goals since then.

I am indebted to my family, first to my older brothers, Yves and Luc, for being so good at school and forcing me to be as good as them and especially to my parents, Hermel and Murielle, who have worked hard to give me the tools to succeed and have always supported me through my various endeavors. Finally, I would like to express my deepest gratitude to Mélanie, for her love, moral and financial support and the many sacrifices she has made since I decided to pursue my dreams, je n'aurais pas réussi sans toi.

Contents

List of Figures

List of Tables

List of Acronyms

ACR	Absolute Category Ratings
AMR	Adaptive Multi-Rate
β-SA	β Spectral Amplitude (β-order STSA MMSE estimator)
dB	decibel
dB SPL	dB Sound Pressure Level
DCR	Degradation Category Ratings
DFT	Discrete Fourier Transform
DTFT	Discrete-Time Fourier Transform
EVRC	Enhanced Variable Rate Codec
GWSA	Generalized Weighted family of STSA estimators
IDFT	Inverse Discrete Fourier Transform
i.i.d.	independent and identically distributed
ITU	International Telecommunication Union
KLT	Karhunen-Loève Transform
LLR	Log-Likelihood Ratio
LSA	Log Spectral Amplitude (MMSE log-STSA estimator)
MAP	Maximum A Posteriori
MMSE	Minimum Mean Square Error
MOS	Mean Opinion Score
MUSHRA	MUlti Stimulus test with Hidden Reference and Anchor
PDF	Probability Density Function
PESQ	Perceptual Evaluation of Speech Quality
SNR	Signal-to-Noise Ratio
SNR_{seg}	Segmental Signal-to-Noise Ratio

STFT	Short-Time Fourier Transform
STSA	Short-Time Spectral Amplitude
VAD	Voice Activity Detector
Wβ-SA	Weighted β Spectral Amplitude (Weighted β-order STSA MMSE estimator)
WCOSH	Weighted COSH
WE	Weighted Euclidian

Chapter 1

Introduction

Speech is one of the predominant means by which humans communicate. The speech signal is generated by the speech production system of a speaker, is transmitted through a certain medium, which can be fiber optic cables, copper wires or simply the air, to finally reach the auditory system of a listener. During this transmission, the speech signal can be corrupted by different types of noises. One of those, additive noise, occurs when an undesired background sound adds itself to the desired speech.

In many common applications it is desirable to suppress such background additive noise. In fact doing so may improve the speech quality, reduce the increased fatigue of the listener caused by the noise or improve the performance of subsequent processing such as that of an automatic speech recognition system or a speech coder. That process of removing a certain amount of background noise in a speech signal is referred to as speech enhancement or more generally as noise reduction.[1] Due to the complex nature of the speech signal, it has been a challenging problem for the past several decades [1–6].

[1]While the term speech enhancement can also be used to designate other aspects such as bandwidth extension of narrow band speech or dereverberation, we will use it here to designate additive background noise reduction.

Below we give a brief overview of the speech enhancement problem and of the different schemes for its solution where we focus on one approach of interest in this thesis, the Bayesian estimation approach. This is followed by a presentation of the different research objectives of this work and the main contributions to the field. Finally, the general organization of the thesis is briefly explained.

1.1 Overview of Bayesian estimators in single-channel speech enhancement

Speech enhancement

The general objective in speech enhancement is to remove a certain amount of noise from a speech signal while keeping the speech components as undistorted as possible. Speech enhancement has been found useful in many applications such as mobile phones [7–9], speech coders [10–12], automatic speech recognition systems [13, 14] and hearing aids [15–18]. In many applications, the noisy speech is acquired by a single microphone. While in some applications, such as hearing aids or hands-free telephony, an array of two or more microphones are sometimes available, we will not consider those cases in this study. Speech enhancement with only one recording of the noisy speech signal is usually qualified as single-channel.

Many single-channel speech enhancement approaches have been proposed over the years. In time domain approaches, which include Kalman filter based methods [19–21], the speech enhancement is performed directly on the time domain noisy speech signal via the application of enhancement filters (i.e. linear convolution). In the frequency domain approaches, a short-time Fourier transform (STFT) is typically applied to a time domain noisy speech signal. The enhancement is then performed by modifying the STFT coefficients which are then

transformed back to the time domain via an inverse STFT. This class of methods includes, among others, spectral subtraction [1–3, 22–27] and Bayesian approaches [4, 6, 28–31]. The enhancement can also be performed in other domains. For example, the so-called subspace approach [32–36] is obtained by applying a Karhunen-Loève Transform (KLT) to the time domain signal, performing the enhancement in that domain and finally going back to the time domain using an inverse KLT operation. In all these approaches, the modification made to the noisy speech depends on the statistical properties of the desired speech and contaminating noise, which must be estimated as part of the enhancement process.

It was found in a subjective comparison of many different speech enhancement methods that the Bayesian approach performed in general better than the other ones [37] in terms of the overall quality of the enhanced speech, the amount of speech distortion introduced by the processing and the background noise reduction. Moreover, compared to other techniques, e.g. the subspace or Kalman-based approaches, their computational requirements are relatively modest. We will concentrate on the Bayesian approach in this thesis.

Bayesian estimation for speech enhancement

In the Bayesian approach for speech enhancement, it is desired to obtain an estimate of the clean speech signal from the noisy speech signal observations. The estimator of the clean speech is derived in the frequency domain by minimizing the expectation of a cost function that penalizes errors in the clean speech estimate. The application of the Bayesian approach therefore requires choosing beforehand a suitable cost function as well as statistical models for the clean speech and noise.

Most existing Bayesian estimators in a frequency domain framework assume the complex STFT coefficients of the clean speech and noise to follow a zero mean complex circular Gaussian distribution with uncorrelated frequency components. The processing is thus

performed on each frequency component independently. The use of a Gaussian statistical model is motivated by the central limit theorem since each Fourier expansion coefficient can be seen as a weighted sum of random variables resulting from the observed samples [4].

One of the simplest cost function that can be used when estimating the clean speech STFT components is possibly the squared error between the estimated and actual clean speech STFT. When combined with Gaussian statistical models with uncorrelated frequency components for the clean speech and noise, this choice results in the widely known Wiener estimator [38, 39]. The Wiener estimator applied to speech enhancement yields fairly good results. However, its main disadvantage is that it produces what are called musical noises. These artifacts are present in the enhanced signal and can be quite annoying to a human listener.

Instead of estimating the STFT coefficients, it is most common to estimate the short-time spectral amplitudes (STSA) and combine them with the phase of the noisy speech to yield an estimate of the STFT. This is justified by the fact that the phase of the STFT coefficients has been shown to be less perceptually significant than the corresponding STSA [40]. One well-known Bayesian STSA estimator is the minimum mean square error (MMSE) of the STSA referred to as the MMSE STSA [4]. It is obtained when the chosen cost function is the squared error between the estimated and actual clean speech STSA and Gaussian models with uncorrelated frequency components are assumed. This estimator was found to produce enhanced speech with a much whiter residual background noise than that of the Wiener estimator. In fact, this property is one of the main advantage of Bayesian STSA estimators for speech enhancement.

Other Bayesian STSA estimators

Many other STSA estimators were proposed over the years. In [28], a more perceptually significant cost function than the squared difference used in the MMES STSA is proposed, where the logarithm of the estimated and actual clean speech STSA are considered. The resulting estimator, which has been termed as log-MMSE STSA (or more conveniently LSA), indeed takes into account that the ear compresses the amplitude of the speech signal.

A generalization of the MMSE STSA cost function was proposed in [6], in which the error between the estimated and actual clean speech STSA is weighted by the STSA of the clean speech raised to an exponent p. Accordingly, the resulting estimator is termed weighted Euclidian (WE). In particular, the author argues that this estimator with $p < 0$ takes advantage of some properties of the human ear and is therefore more perceptually significant.

Another generalization of the MMSE STSA cost function is also proposed in the β-Order STSA MMSE estimator (which we will denote by β-SA for convenience) [29]. This estimator applies a power law (i.e. an exponent $\beta > 0$) to the estimated and actual clean speech STSA in the squared error of the cost function. As a particular case, the authors observed through numerical calculation that in the limit $\beta \to 0$, the β-SA estimator approaches the LSA estimator.

The values of the parameter β in the β-SA estimator and p in the WE estimator were found to control the trade-off between the speech distortion introduced by the enhancement and the background noise reduction. In fact, in single-channel speech enhancement, reducing the background noise level will invariably produce some speech distortions and a trade-off between the desired noise reduction and the undesired speech distortions must be achieved.

While many speech enhancement estimators have been proposed in the past decades, the existing estimators are far from ideal and suffer from many problems including speech distortions at various degrees and some residual noise artifacts. There is thus still much space for improvement in order to further remove the background noise while limiting the speech distortions, and indeed there has been significant research efforts in this direction in recent years.

1.2 Research objectives

The general objective of this research work is to study Bayesian STSA estimators for single-channel speech enhancement in order to deepen the existing knowledge on such estimators and improve their performance. This is to be achieved by analyzing existing Bayesian STSA estimators and proposing more appropriate cost functions and statistical models.

In light of Section 1.1, many improvements over traditional approaches are conceivable. In particular, we study the following questions in this thesis:

- The starting point of our work is to study the β-SA estimator and the role played by its parameter β. For example, only the values of $\beta > 0$ are considered in [29]; however, it can be shown that the derivation of the estimator allows for values of $\beta > -2$. These negative values of β may reveal some advantages in terms of the quality of the corresponding enhanced speech over their positive counterparts and suggest new modifications to the underlying cost function.

- Some parameters such as the exponent β in the β-SA estimator and p in the WE estimator have auditory or perceptual significations that have not been considered yet. As discussed in the previous section, the human auditory system can provide some useful insight that can be used to develop speech enhancement estimators.

Choosing the values of those parameters based on such considerations could result in more perceptually significant cost functions and better estimators. One particular objective of our work will therefore be to study values for such parameters that have auditory or perceptual significations and to evaluate their effect on the resulting speech enhancement estimators.

- Most existing Bayesian STSA estimators, including those based on the WE and β-SA cost functions, have a very similar structure consisting of a weighted squared difference between a monotonic function of the estimated and actual clean speech STSA. One particular objective of this work is to combine some of these cost functions into a more general framework for the derivation of Bayesian STSA estimators; this should allow a better understanding of this class of estimators.

- Finally, it is always assumed in single-channel Bayesian STSA estimators that the frequency components of the clean speech and noise are not correlated. Therefore, each frequency component is treated independently. In practice, however, some of the frequency components are correlated for reasons explained later and cannot be treated independently. Improvements in the quality of the enhanced speech could be obtained by considering this correlation in the cost function and statistical models. Related objectives of this work are thus to evaluate how such correlation can be included in clean speech models and develop new Bayesian STSA estimators that exploit this correlation.

1.3 Main contributions

In this thesis, by pursuing the objectives set forth in the previous section, we extend the knowledge on Bayesian STSA speech enhancement estimators and propose several new

estimators that show some advantages over existing ones. The main original contributions of this research work are summarized below. We first present the different analytical contributions which are followed by a summary of the experimental results.

Analysis and extension of the β-SA estimator

We first show that negative values of β have a normalization effect on the original β-SA cost function. Moreover, decreasing β below 0 is found to produce an increase in the noise reduction and speech distortion, therefore enabling an extension of the trade-off between speech distortion and noise reduction. We also observe, based on gain curves, that the β-SA estimator with $\beta < 0$ behaves similarly to the WE estimator [6] for negative values of p. Finally, it is proved analytically that in the limit case $\beta \to 0$, the β-SA estimator indeed corresponds to the LSA estimator of [28].

The weighted β-SA estimator and perceptually relevant parameter values

A new Bayesian STSA estimator, that we call the weighted β-SA (Wβ-SA) estimator, is developed and investigated. This estimator combines the power law in the β-SA cost function and the weighting of the WE cost function to take advantage of the perceptual interpretation that can be given to the associated parameters. The values of the parameters in the corresponding cost function, and therefore of the corresponding estimator gain, are chosen based on characteristics of the human auditory system namely loudness, masking and the compressive nonlinearities of the ear. It is found that doing so suggests a decrease in the gain at high frequencies. This decrease in the gain has the advantage of limiting the speech distortions at low frequencies, where the main speech energy is located, while increasing the noise reduction at high frequencies.

Analytical generalization of Bayesian STSA estimators

We show that many existing Bayesian STSA estimators for speech enhancement all have a similarly structured cost function. On this basis, an analytical generalization of Bayesian STSA estimators is performed where: (1) a new cost function that unifies existing Bayesian STSA estimators is proposed and; (2) the corresponding closed-form solution for the optimal clean speech STSA is obtained. The resulting family of estimators, which we term the Generalized Weighted family of STSA estimators (GWSA), includes many existing estimators as particular cases and approaches a Wiener filter in the limit of high instantaneous SNR. It also features a new parameter that acts only on the estimated clean speech STSA. It is found that this new parameter yields an added flexibility in terms of achievable gain curves when compared to those of existing estimators.

Multi-dimensional extension of Bayesian STSA estimators

In Bayesian STSA estimation for single-channel speech enhancement, the spectral components are traditionally assumed uncorrelated. However, this assumption is inexact and we show, in fact, that there is some correlation between STFT speech coefficients. We therefore investigate a multi-dimensional Bayesian STSA estimator that assumes correlated frequency components. Since the closed-form solution of this optimum estimator is not readily available, we alternatively derive simple closed-form expressions for an upper and a lower bound on the desired estimator. Based on these bounds, we finally propose a family of speech enhancement estimators that are characterized by a one-dimensional parameter $0 \leq \gamma \leq 1$ with $\gamma = 0$ corresponding to the lower bound and $\gamma = 1$ to the upper bound.

Experimental validation

In this thesis, the validity of the proposed estimators is demonstrated experimentally. To do so, they are compared with relevant state-of-the art estimators including Wiener [3, 41], MMSE STSA [4], LSA [28], WE [6] and β-SA [29]. This is achieved through the use of several objective and subjective measures. The former include the segmental SNR (SNR_{seg}) [42], log-likelihood ratio (LLR) [43] and perceptual evaluation of speech quality (PESQ) [44] while the subjective ones include informal listening tests, mean opinion scores (MOS) [45] and the multi-stimulus test with hidden reference and anchor (MUSHRA) [46].

The β-SA with negative values is first evaluated. It is found that with $\beta = -1$, its performance is slightly better than the MMSE STSA and LSA estimators in terms of the PESQ and that the overall MOS appreciation of the β-SA with $\beta = -1$ is better than both MMSE STSA and LSA. Secondly, experimental results show that the Wβ-SA estimator with the proposed frequency dependent parameter values achieve better enhancement performance than all compared estimators (i.e. MMSE STSA, LSA and WE ($p = -1$)) in terms of all studied measures i.e. SNR_{seg}, LLR, the wideband extension of PESQ and MUSHRA. This advantage of the proposed estimator is shown to be substantial at low SNR values. We finally compare the proposed family of multi-dimensional estimators that considers correlated STFT components with the traditional Wiener and MMSE STSA estimators (i.e. which consider uncorrelated frequency components) as well as with an MMSE estimator of the complex STFT coefficients that assumes correlated frequency components. Informal listening experiments as well as results using the wideband PESQ and LLR measures all show that the proposed family of multi-dimensional estimator achieves better performance than the benchmark estimators for several noise types and SNR conditions.

Publications

These contributions led to a number of publications in peer-reviewed journals and refereed conferences. The following is a list from this thesis work:

Journal papers

J-1 E. Plourde and B. Champagne, "Multi-dimensional Bayesian STSA estimators for the enhancement of speech with correlated frequency components," *IEEE Transactions on Audio, Speech and Language Processing*, submitted, August 2009.

J-2 E. Plourde and B. Champagne, "Generalized Bayesian estimators of the spectral amplitude for speech enhancement," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 485-488, June 2009.

J-3 E. Plourde and B. Champagne, "Auditory based spectral amplitude estimators for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1614-1623, November 2008.

Conference papers

C-1 E. Plourde and B. Champagne, "Bayesian spectral amplitude estimation for speech enhancement with correlated frequencies," *IEEE Workshop on Statistical Signal Processing*, August 31 - September 3, 2009, Cardiff, Wales, UK.

C-2 E. Plourde and B. Champagne, "Perceptually based speech enhancement using the weighted β-SA estimator," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, March 30 - April 4, 2008, Las Vegas, NV, USA, pp. 4193-4196.

C-3 E. Plourde and B. Champagne, "Integrating the cochlea's compressive nonlinearity in the Bayesian approach for speech enhancement," in *Proc. 15th European Signal Processing Conf. (EUSIPCO)*, September 3-7, 2007, Poznan, Poland, pp. 70-74.

C-4 E. Plourde and B. Champagne, "Further analysis of the β-Order MMSE STSA estimator for speech enhancement," in *Proc. 20th IEEE Canadian Conf. on Electrical and Computer Eng. (CCECE)*, April 22-26, 2007, Vancouver, BC, Canada, pp. 1594-1597.

1.4 Thesis organization

Chapters 2 and 3 present background material while the following chapters present the different contributions.

In Chapter 2, some fundamentals of human speech production and the auditory system are reviewed and important aspects of noise contamination in speech signals are discussed. An overview of the Bayesian approach for single-channel speech enhancement along with different existing estimators are presented in Chapter 3.

In Chapter 4, we analyze the β-SA estimator for $\beta < 0$ and also prove analytically that for $\beta \to 0$, the β-SA estimator is equivalent to the LSA estimator. The Wβ-SA estimator is introduced in Chapter 5 where we also propose frequency dependent values for its parameters that are based on auditory and perceptual considerations. In Chapter 6, we perform an analytical generalization of several existing estimators. A framework that considers correlated frequency components in Bayesian STSA estimation is presented in Chapter 7 along with the proposed family of new estimators.

Finally, in Chapter 8, we present experimental results for the proposed estimators. Some concluding remarks are presented in Chapter 9.

Chapter 2

Human speech communication

From a signal processing perspective, a speech communication between two or more individuals is a highly complex process that involves several elements. First, the speech signal is produced by the speech production system of the individual who is speaking. This signal is then transmitted through a certain medium and during its transmission, is subjected to several forms of modification and noise contamination. This contaminated signal then arrives at the ears of the other individuals and is processed by their auditory systems.

In Sections 2.1 and 2.2 of this chapter, respectively, we will briefly look at both the human speech production and auditory systems. Section 2.3 will present possible sources of speech contamination and discuss applications where it can be useful to remove the added noise.

2.1 Human speech production system

The speech production system is divided into the organs of phonation and articulation. The phonatory organs are the lungs and the larynx which includes the vocal cords. They

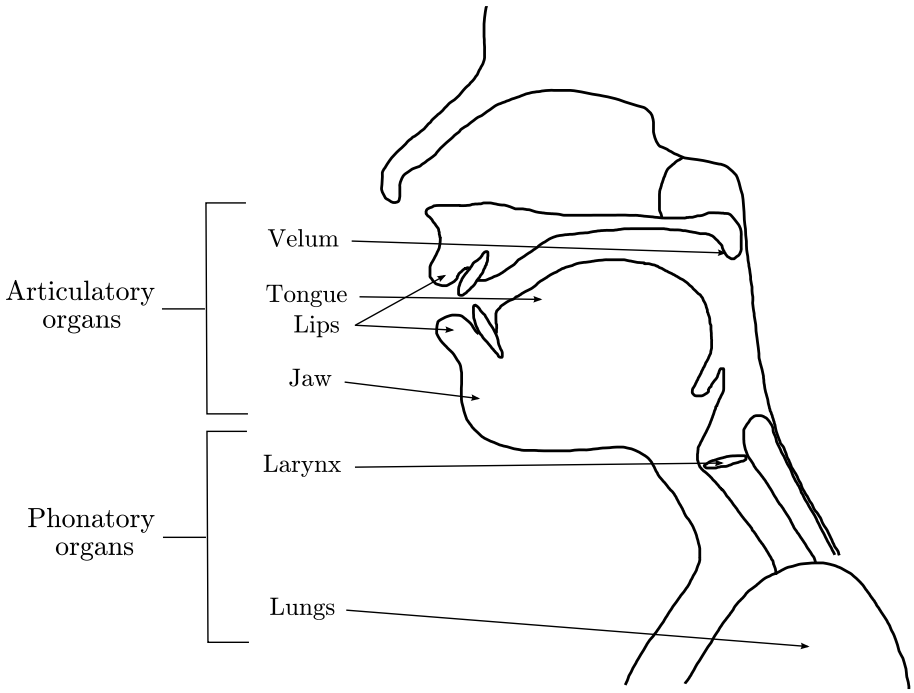

Fig. 2.1 The human speech production system (inspired from [47, 48]).

create the voice source sounds by initiating the exhaled air pressure and controlling the vocal cord vibrations. These organs adjust the pitch, loudness and quality of the voice. The articulatory organs (i.e. jaw, tongue, lips and velum) modulate the voice source sound and generate some consonants (see Fig. 2.1). The different positioning of those articulatory organs coupled with the sound source will allow the speaker to produce different phonemes, i.e. to speak.

Speech production can be modeled by a sound source representing the lung and the larynx that excites a so-called vocal tract filter where the vocal tract is defined as the oral cavity from the larynx to the lips and the nasal passage that is coupled to the oral tract by way of the velum [49]. If the vocal cords vibrate, the source is then periodic and the speech is qualified as voiced, while if the vocal cord is constricted, the source is noisy and the speech is qualified as unvoiced [50]. The vocal tract filter amplifies or attenuates certain sound

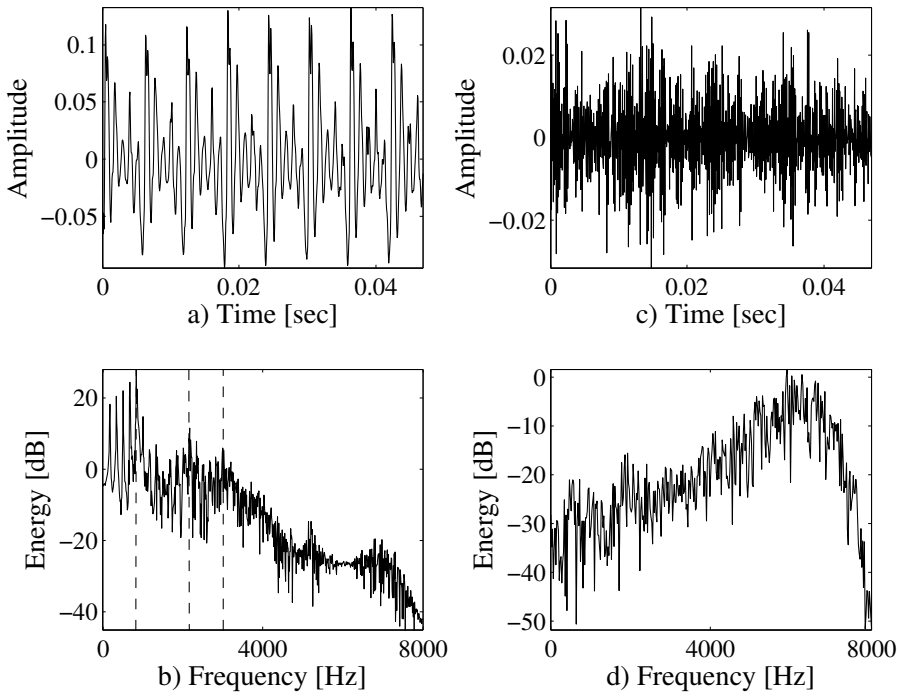

Fig. 2.2 (a) Time domain female uttered voiced phoneme (/ε/); (b) Frequency domain representation of (a); (c) Time domain female uttered unvoiced phoneme (/s/); (d) Frequency domain representation of (c).

frequencies for both voiced and unvoiced source sounds depending on the positioning of the articulatory organs. The different resonating frequencies of the vocal tract are termed formants and are labeled as F1, F2, F3, etc. The smallest number indicates a smaller formant frequency and vice versa.

Speech sounds are classified into different phoneme classes. Among them are vowels, semi-vowels, consonants, affricates and diphthongs . Vowels are voiced sounds for which the vocal tract configuration is kept fixed. They are characterized by the value of their formant frequencies. Consonants can be voiced or unvoiced and include a number of subgroups of phonemes such as nasals, plosives, fricatives and whispers. Finally, the semi-vowels (including the liquids and glides subgroups), affricates and diphthongs are transitional speech sounds [49].

Fig. 2.2 shows the waveform and energy spectrum of a voiced (/ε/) and an unvoiced (/s/) speech phoneme uttered by a female speaker. As shown in Fig. 2.2 (a), the voiced phoneme is periodic in the time domain due to the vocal cord vibrations. Moreover, voiced speech has a spectra that consists of harmonics of the fundamental frequency of the vocal cord vibrations. That fundamental frequency is usually termed F0 and corresponds to the perceived pitch. We can observe in Fig. 2.2 (b), particularly at the lower frequencies, the different harmonics separated by F0 $\approx$ 160 Hz for this speaker. We can also identify the formant frequencies F1, F2 and F3 at approximately 800 Hz, 2100 Hz and 3000 Hz respectively. They correspond to the resonating frequencies of the vocal tract. The frequencies where formants are located are usually termed spectral peaks while the regions in between formants are termed spectral valleys. Finally, we notice that the energy of voiced speech is mostly concentrated at the lower frequencies.

As opposed to voiced speech, we can see from Fig. 2.2 (c) that unvoiced speech does not show a periodic structure but has a noisy nature. Moreover, we can also observe from Fig. 2.2 (d) that the energy of unvoiced speech is much less than that of voiced speech and mostly concentrated at higher frequencies.

2.2 Human auditory system

2.2.1 The peripheral auditory system

At the other end of the human communication system is the auditory system that captures and interpret the speech signal. The peripheral auditory system, commonly termed the ear, is composed of three main parts (Figure 2.3):

- Outer ear: composed of the pinna and the ear canal or meatus;

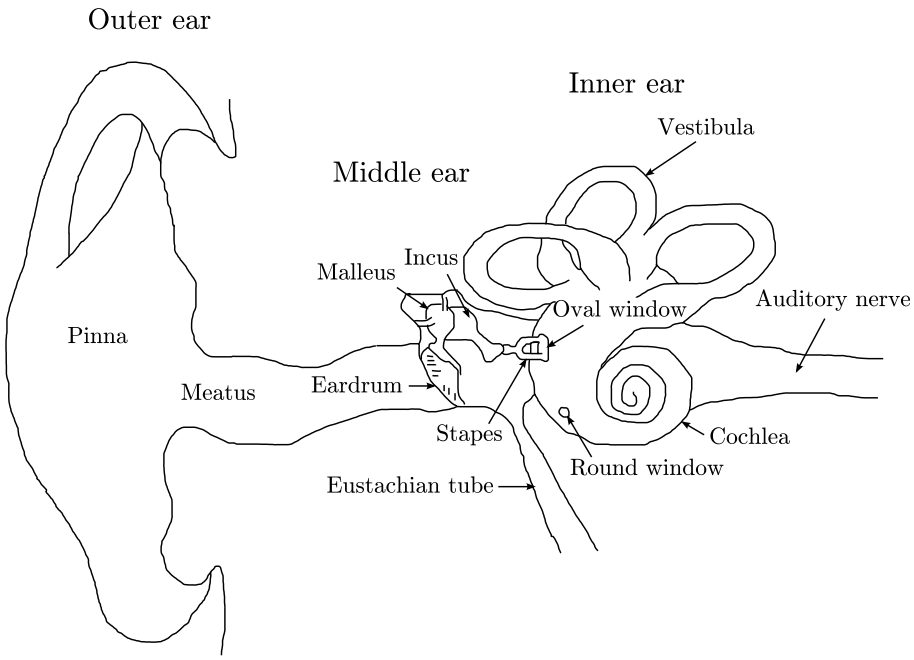

Fig. 2.3 The structure of the peripheral auditory system (inspired from [50]).

- Middle ear: composed of the eardrum, malleus, incus and stapes;
- Inner ear: composed of the cochlea.

A sound wave first enters the ear through the pinna and progresses through the meatus. Once it reaches the eardrum, the sound wave produces vibrations of the middle ear bones (malleus, incus and stapes) which together act as a transducer between the air medium of the ear canal and the liquid medium of the cochlea. The middle ear bone vibrations are then carried in the cochlea through the oval window. The cochlea transforms the vibrations into neuronal impulses which are then carried for further processing in the brain [51].

The cochlea

The cochlea plays an important role in auditory speech processing. It is a fluid-filled tube which is coiled in a snail-shaped spiral and has a total diameter of approximately 9 mm [51].

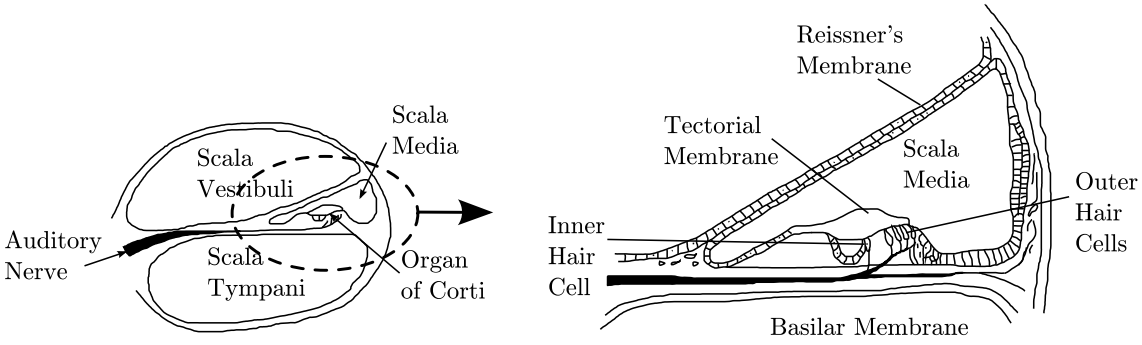

Fig. 2.4 Structural and anatomical features of the cochlea and basilar membrane (inspired from [50]).

The tube is separated by two membranes of which the basilar membrane has a fairly important role. When a sound wave hits the oval window, it creates a pressure difference between both sides of the basilar membrane which makes it oscillates at different locations depending on the incoming sound frequency.[1] In effect, the basilar membrane performs a frequency analysis of the incoming sound wave. Regions close to the oval window, called the base of the cochlea, oscillates at higher frequencies while regions towards the end of the basilar membrane, the apex, oscillates mainly at lower frequencies. The movement of the basilar membrane is sensed by sensory cells, the hair cells, which activate the firing of the neurons. Figure 2.4 shows a cross section of the cochlea along with a zoom on the basilar membrane. As shown, the hair cells, which are of two forms, i.e. inner hair cells and outer hair cells, are part of the organ of Corti. The inner hair cells are primarily responsible for the transduction of the basilar membrane movement into neuronal firing. The outer hair cells, on the other hand, play an important role in cochlear amplification and are responsible for the active mechanism of the cochlea [53].

[1] The organization where a position on the basilar membrane corresponds to an associated frequency is termed *tonotopic* [52].

Auditory neurons

The inner hair cells are connected to the auditory nerve which relays the sound information to the ascending neural pathway in the brain. The auditory nerve is composed of many neurons each able to sustain a maximum firing rate of 500 spikes per second in response to bending of the hair cells. This bending of the hair cells alters their electrical conductance which influences the release of a chemical substance (neurotransmitter) which causes the corresponding neuron to fire [51].

2.2.2 Relevant properties of the auditory system

The human ear has a remarkable ability to select a desired speech signal among a noisy background. While doing so, it naturally suppresses a part of the background noise. In this subsection, we will briefly present some properties of the auditory system that will be particularly useful in some sections of this thesis. In particular, we will consider auditory masking and the compressive nonlinearity of the cochlea.

Auditory masking

Auditory masking is the process by which the threshold of audibility of a particular signal is raised by the presence of another sound (the masker) [54]. Therefore, a noise may naturally be masked by a speech component, and therefore not be heard. Masking is thought to have its origins on the basilar membrane and may arise in the process of hair cell firings [50]. Two distinct types of masking actually occurs: simultaneous masking and temporal masking.

On the one hand, simultaneous masking occurs when the masker is present during the presentation time of the signal of interest. Masking effects are frequently taken into account through masking thresholds below which a tone will be masked and therefore not

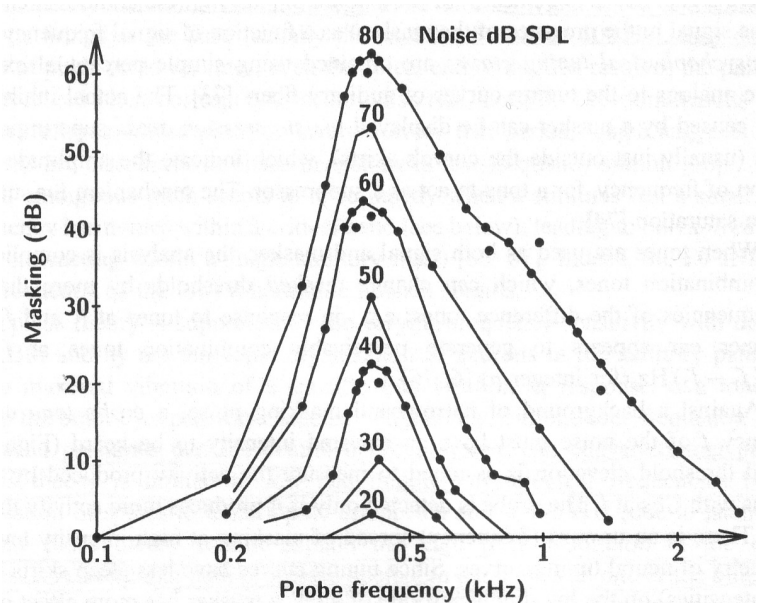

Fig. 2.5 Simultaneous masking thresholds. (from [50])

be heard. Figure 2.5 illustrates the masking thresholds in simultaneous masking produced by a narrow band of noise (365-455 Hz) for different intensities. Among other things, one notices that there is an asymmetry in simultaneous masking for higher intensities (i.e. 80 dB SPL) which disappears for lower intensities.

Temporal masking, on the other hand, occurs when the signal is presented just before or after the masker [54]. In fact, successive signals with energy in the same critical band[2] can interfere with one another if the delay between them is sufficiently short. Temporal masking can be of two kinds: forward masking where for example, a time limited noise signal can mask a following tone, and backward masking, where the noise masks a preceding tone. Forward masking can exist for a delay between the noise and the tone of as much as 200 ms; however, backward masking has effects only when the delay is below 20 ms [50].

[2]The concept of critical bands was proposed by Fletcher who assumed that the part of a noise that is effective in masking a test tone is the part of its spectrum lying near the tone, i.e. in a surrounding frequency band called a critical band. The relative powers of the noise and the tone in the entire critical band determines if the tone is masked or not. [55]

Compressive nonlinearity of the cochlea

The ear is most sensitive to small signals and grows progressively less responsive as stimulation becomes stronger. This allows us to interpret sounds over a wider range of amplitudes and is also thought to play a role in the noise suppression capabilities of the auditory system [56]. This nonlinearity appears on the basilar membrane and is thought to be frequency dependent.

A - High frequencies (base of the cochlea): Researchers have noticed a nonlinear behavior at the base of the cochlea, which is associated to the processing of high frequencies, when measuring basilar membrane responses to input tones at several sound pressure levels. Figure 2.6 shows measurements of the basilar membrane displacement versus the input sound pressure levels performed on several mammals, the characteristic frequencies of the tones varied from 8 kHz to 33 kHz depending on the species. Equivalent data for humans is not available other than from cadavers which are known not to retain the compressive response. However, the existence of a similar behavior in the human auditory system has been indirectly confirmed through various psychoacoustic measures [57]. Looking at Fig. 2.6, one notices that compared to the linear growth rate, the basilar membrane exhibits a so-called compressive nonlinearity[3]. In fact, for high input sound pressure level, the output is compressed where as for lower levels, the output may be expanded or amplified. We will use the term compressive nonlinearity to denote the overall phenomenon.

This nonlinearity is thought to be caused by the active mechanism of the outer hair cells which at lower input amplitudes exhibit an amplification of the basilar membrane vibration, termed cochlear amplification. As the amplitude increases, however, this amplification

[3]The growth rate, or compression rate, is defined as the slope of a displacement versus sound pressure level curve where the displacement is expressed in dB. A compression rate of 1 indicates a linear relationship and therefore no compression.

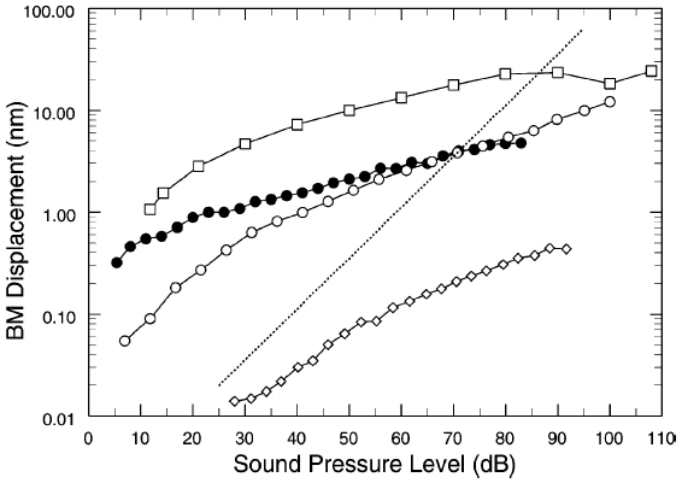

Fig. 2.6 Displacement of basilar membrane versus sound pressure level for characteristic frequency tones recorded at basal cochlear sites in chinchilla (□), guinea pig(○, ●), and cat (◇) – For comparison, the dotted line indicates linear growth (from [53]).

saturates and, in relative terms, the larger spectral amplitudes become compressed. Compression rates of 0.2 dB/dB were measured for intensities between 40 and 90 dB SPL [53] (conversational speech is at 60 dB SPL) and they tend to be more linear, i.e. closer to 1 dB/dB, for lower intensities.

B - Low frequencies (apex of the cochlea): The apex of the cochlea is associated to the processing of low frequencies. While compressive nonlinearity is well documented and accepted for high frequency signals, there is no real consensus on the degree of cochlear compressive nonlinearity at lower frequencies. In fact, some results from chinchilla show a small rate of compression (0.5 - 0.8 dB/dB), while several other results from guinea pigs and squirrel monkeys fail to show any compressive nonlinearity (i.e. they reported a rate of compression of 1 dB/dB) and even show an expansion (i.e. rate of compression greater than 1 dB/dB) [53]. Besides, psychoacoustic experiments in humans report either a comparable rate of compression at low and high frequencies [58,59] or a smaller but existent rate of compression at lower frequencies [60]. Since those results are from psychoacoustic

experiments and not from a specific physiological experiment, one cannot be sure where in the auditory processing path this compression occurs and it may not be a cochlear phenomenon but rather a property occurring along the auditory neural pathway [58, 59]. Therefore, it is usually assumed that there is a difference in the cochlear rate of compression at high and low frequencies [53], but its relative values are still an active debate.

2.3 Contamination in speech signals

While a speech signal would ideally leave the speakers lips and be transmitted unaltered to the listener's ears, in reality, this signal is modified by different forms of contamination during its transmission. Two basic types of contamination can be identified, namely: convolutive and additive noises.

Reverberation, or convolutive noise, arises when the speech signal propagates through multiple paths with different transmission times and therefore arrives in delayed versions at the listener's ears. In this case, the received signal can be expressed as a mathematical convolution between the clean source signal and the unknown transmission impulse response. The perceived reverberation is particularly important when the ratio of the direct signal to the echoed signal is small. In this thesis, we consider this ratio to be large enough and concentrate only on additive noise.

Additive noise occurs when one or more other sound sources are added to the desired speech signal as the result of a linear superposition in the acoustic medium. The other sound source can be, for example, a nearby speaker, a car passing by on the street or background music.

Additive noise contamination can be undesirable for several reasons. In fact, it can affect, for example, different perceptual aspects of the speech signal such as its intelligibility

and quality. Intelligibility refers to the number of words that can be identified correctly by a listener or to the likelihood of being correctly understood whereas quality refers to the clarity, freedom of distortion and ease for listening [50]. The two are not correlated such that a good quality speech signal can have a poor intelligibility [61]. Apart from affecting perceptual aspects of speech, noise contamination can also affect the performance of speech processing applications such as speech coders, teleconferencing or automatic speech recognition. In many common applications, there is thus a motivation to remove such undesired noise, we discuss some of these applications below.

A - Wireless telephony: One important application in which speech is subjected to additive noise contamination is cellular or wireless telephony. In fact, with the advent of portable phones, the environment in which a telephone communication occurs went from the traditional house or office setting to much diverse environments such as crowded streets, cars, public transportations, restaurants, etc. These environments can be characterized by much lower signal-to-noise (SNR) ratios than traditional environments and the quality and intelligibility of the speech is sometimes greatly diminished by the noise contamination. It is therefore highly desirable to avoid such degradation, and in fact, much research has been done to achieve noise reduction in mobile phones [7, 8]. Moreover, speech codecs used in mobile phones are generally less efficient in the presence of noise. In fact, many codecs, such as the Enhanced Variable Rate Codec (EVRC) [62] or the codec defined by the G.711.1 standard [63], integrate some noise reduction modules.

B - Teleconferencing: Teleconferencing is typically a hands free application. It allows many persons in a room to interact with one or many other groups of persons in a different physical setting through the use of one or more microphones and loudspeakers. Due to its hands free characteristic, the different listeners in a teleconference will be subject to any ambient noise entering the system. Efforts have thus been made to remove that noise [64].

C - Automatic speech recognition: In the past decade, various automatic speech recognition systems have been incorporated in different applications such as hands free cellphones or as a substitute to phone operators. These systems are also affected by noise and, in fact, their performance may decrease significantly when a noisy instead of a clean speech signal is used as an input to the system. The reduction of noise in such systems has thus also fostered much research [13, 14].

D - Hearing aids: Finally, another application where the reduction of background noise is useful is in hearing aids. Indeed, persons with hearing deficiencies are generally more affected by noise than normal hearing persons. This is due in part to their resolution of the different spectral components of speech which is not as efficient as that of a normal hearing person. Hearing impaired are thus less capable of discerning noise from speech. Research has therefore been performed to incorporate in hearing aids some speech enhancement modules reducing the effect of the noise contamination [15–18].

The next chapter will look more in detail at how this noise reduction can be achieved. In particular, we will present the speech processing used in single-channel speech enhancement and the corresponding existing estimators, mainly Bayesian estimators.

Chapter 3

Overview of the Bayesian approach for single-channel speech enhancement

As discussed in the previous chapter, there is a strong motivation in many applications to remove the additive noise contaminating a speech signal. In order to do so, many speech enhancement approaches have been derived over the years. In this chapter, we will start by presenting the single-channel speech enhancement problem, set in the frequency domain. Speech enhancement algorithms using the Bayesian approach in a frequency domain framework were recently found to give the best results among various competing approaches including subspace and spectral subtraction [37]. We will thus review several algorithms using the Bayesian approach for single-channel speech enhancement. In particular, we will present Bayesian estimators of the STFT coefficients as well as Bayesian estimators of the STSA.

This chapter is organized as follows. In Section 3.1, we present the speech enhancement

problem in the frequency domain. In Section 3.2, the Bayesian estimation framework is developed and in Section 3.3 and 3.4, we present Bayesian estimators of the STFT and STSA respectively. Finally, in Section 3.5, we elaborate on the estimation of the *a priori* and *a posteriori* SNR parameters.

3.1 The single-channel speech enhancement problem

3.1.1 Additive noise model

In some applications, such as hearing aids and multi-channel teleconferencing, several microphones can be used and therefore many versions of the noisy speech are available simultaneously[1]. However, to limit the system's constraints, only one microphone is present in the majority of applications. The process of removing noise when only one source of the noisy speech is available is referred to as single-channel speech enhancement. Single-channel speech enhancement algorithms are generally not able to improve the intelligibility of speech, discussed in the previous chapter, but improve the quality of speech [65].

As mentioned previously, many types of noise can contaminate a speech signal such as convolutive or additive noises. For reasons explained in Chapter 2, we will concentrate on additive noises, such as ambient background noise, in this thesis. Let an observed noisy speech signal be represented by the following additive noise model as illustrated in Fig. 3.1:

$$y[m] = x[m] + w[m] \qquad 0 \leq m < L \tag{3.1}$$

where $x[m]$ represents the unknown clean speech, $w[m]$ is the additive noise, m is the discrete-time index, where uniform sampling is assumed, and L is the total number of

[1]Some manufacturers are actually offering mobile phones with two microphones. They are used to perform a basic beamforming function, after which single channel speech enhancement can be applied.

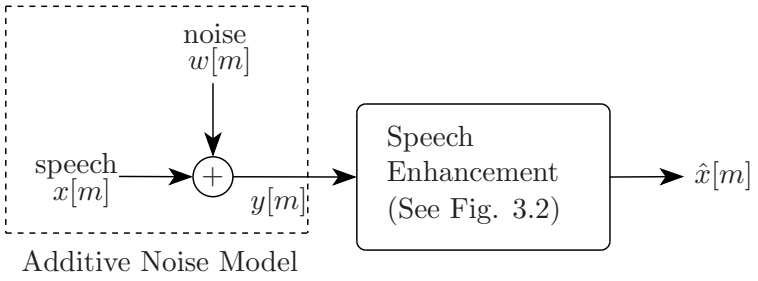

Fig. 3.1 Additive noise model in single-channel speech enhancement.

observed samples. The objective in single-channel speech enhancement algorithms is to find an estimate $\hat{x}[m]$ of $x[m]$ from the noisy speech observations $y[m]$ as shown in Fig. 3.1.

Two main processing frameworks can be used to obtain the desired estimate: time domain estimation, where the enhancement is performed directly on $y[m]$ to yield $\hat{x}[m]$; or frequency domain estimation where the enhancement is performed on the Discrete Fourier Transform (DFT) coefficients obtained from $y[m]$. According to [66], the computational demand is less for frequency domain algorithms than for time domain algorithms, which makes the former more attractive for low-power applications. We will consider only the frequency domain framework in the sequel.

3.1.2 Frequency domain single-channel speech enhancement

We now look in more detail at the frequency domain framework used in single channel speech enhancement algorithms. Within a short observation interval of about 20-40 ms, a speech signal $x[m]$ is generally considered to be a realization of a zero mean and wide-sense stationary random process. It is therefore of common practice to separate the set of L observed samples into overlapping frames lasting less than 40 ms[2]:

$$y_i[n] = y[n + iM] \quad 0 \leq n < N, \quad 0 \leq i < N_f \tag{3.2}$$

[2]For example, at a sampling rate of 16 kHz, a 32 ms frame would correspond to $N = 512$ samples.

where i denotes the frame index, M is the frame advance, N is the frame length with $N \geq M$ ($N - M$ is the number of samples that overlap between two successive frames) and N_f is the total number of frames in the L observed samples. An analysis window $h_a[n]$ is applied on each frame in order to achieve a trade-off between frequency resolution and sidelobe suppression [67]. Each windowed frame is then transformed in the frequency domain using a discrete Fourier transform (DFT):

$$Y_{k,i} = \sum_{n=0}^{N-1} y_i[n] h_a[n] e^{-j\frac{2\pi}{N}kn} \tag{3.3}$$

where $k \in \{0, 1, \ldots, N-1\}$ denotes the frequency index[3]. $Y_{k,i}$ is referred to as the k^{th} short-time Fourier transform (STFT) coefficient of the noisy speech for the i^{th} frame [49]. With $X_{k,i}$ and $W_{k,i}$ denoting in the same way the STFT coefficients of the clean speech and noise respectively, the additive noise model in the STFT domain thus becomes:

$$Y_{k,i} = X_{k,i} + W_{k,i}. \tag{3.4}$$

In the STFT domain, the objective is therefore to find an estimate $\hat{X}_{k,i}$ of $X_{k,i}$ from $Y_{k,i}$. Once $\hat{X}_{k,i}$ is obtained, its time-domain counterpart $\hat{x}_i[n]$ is derived by applying an inverse DFT (IDFT) on each frame:

$$\hat{x}_i[n] = \frac{1}{N} \left(\sum_{k=0}^{N-1} \hat{X}_{k,i} e^{j\frac{2\pi}{N}kn} \right) h_s[n] \tag{3.5}$$

where $h_s[n]$ is a proper synthesis window.

[3] N is often chosen as the nearest power of 2.

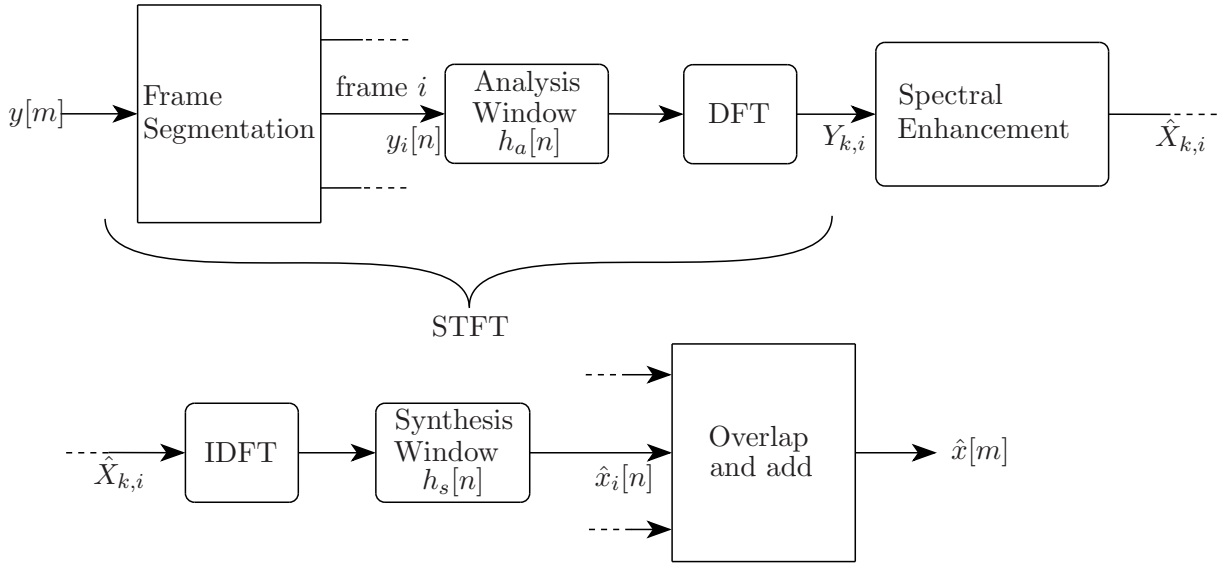

Fig. 3.2 STFT framework for single-channel speech enhancement.

The results for all frames are then combined using an overlap-add reconstruction [67,68]:

$$\hat{x}[m] = \sum_{i=0}^{N_f-1} \hat{x}_i[m - iM]. \tag{3.6}$$

The overall process of speech enhancement in the STFT domain is illustrated in Fig. 3.2. In practice, since the signals under consideration are real valued, the DFT coefficients from $k = N/2+1$ to $k = N-1$ are the complex conjugates of the coefficients from $k = N/2-1$ to $k = 1$ respectively. Furthermore, the different processes are generally considered to be zero-mean and properly bandlimited, so that the DFT coefficients corresponding to $k = 0$ and $k = N/2$ can be taken as 0. Therefore, the different signals can be processed only for frequencies $k = 1$ to $k = N/2-1$.

To ensure that the combined analysis and synthesis windowing process does not introduce unwanted modifications to the speech signal, the analysis and synthesis windows need

to satisfy the following perfect reconstruction condition [67]:

$$\sum_{i=0}^{N_f-1} h_a[m-iM]h_s[m-iM] = 1 \quad \forall m. \tag{3.7}$$

In this thesis, this condition will be met by choosing $h_a[n]$ as a raised-cosine window and $h_s[n]$ as a rectangular window, i.e.

$$h_a[n] = \begin{cases} \frac{1}{2} + \frac{1}{2}\sin\left(\frac{\pi}{M'+1}\left(n - \frac{M'-1}{2}\right)\right) & 0 \le n < M' \\ 1 & M' \le n < M \\ \frac{1}{2} - \frac{1}{2}\sin\left(\frac{\pi}{M'+1}\left(n - M - \frac{M'-1}{2}\right)\right) & M \le n < M + M' \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

$$h_s[n] = 1, \quad 0 \le n < N \tag{3.9}$$

where the length of the tapered window is $M + M'$.

According to [69], any linear modification to the signal $Y_{k,i}$ within the STFT framework of Fig. 3.2, i.e. $\hat{X}_{k,i} = G_k Y_{k,i}$ where G_k is a complex gain, is equivalent to linear filtering in the time domain via circular convolution. This filtering increases the length of the time domain signal and can result in time aliasing due to its circular nature. To limit time aliasing, we can increase the length of the window $h_a[n]$ by appending some $N_z - 1$ zeros to it in which case we have $N = M + M' + N_z - 1$. This is referred to as zero-padding and increases the length of the DFT [69].

The *Spectral Enhancement* block in Fig. 3.2 can be realized by means of several frequency domain approaches among which the best known are the spectral subtraction, Wiener and Bayesian STSA approaches.

Spectral subtraction [1–3, 22–27] has been intensively studied over the past forty years.

It attempts to estimate the spectral amplitude of the clean speech by subtracting an estimate of the noise spectral amplitude from that of the observed noisy speech. Different means are used to ensure that the spectral amplitude estimator resulting from the subtraction does not have a negative value. Finally, the estimated amplitude is combined with the phase of the noisy speech to produce the desired estimate of the clean speech STFT. In the power spectral subtraction variant, an estimate of the spectral amplitude of the clean speech is obtained by subtracting an estimate of the noise power spectrum from that of the noisy speech and applying a square root to the resulting estimator. The phase is handled in the same way as in the spectral subtraction scheme.

The major drawback in spectral subtraction techniques is the musical nature of the residual noise. Musical noises are characterized by tones at different frequencies that randomly appear and disappear. They can be extremely annoying to a human listener. Many versions of the spectral subtraction approach have been proposed over the years [1–3,24,25,27], most of which attempted to reduce the amount of musical noise in the resulting estimator.

In the Wiener filter approach [3,41,70,71], which will be discussed more in detail in the next subsection, the estimator of the clean speech STFT is simply the MMSE estimator when considering Gaussian distributed clean speech and noise. In that case, the phase of the resulting estimate turns out to be that of the noisy speech. As for the spectral subtraction approach, the speech enhanced based on the Wiener filter is also characterized by residual musical noises.

We will concentrate on the Bayesian STSA approach [4,6,28,30,36,39,72–76] in this thesis which has the strong advantage of producing mostly white residual noise. The Bayesian STSA approach along with recent related developments are explained in detail in the following section. The interested reader may consult [48,77] for a review of alternative speech enhancement approaches, including the above spectral subtraction and Wiener filtering as

well as other approaches such as the subspace [32–35] or Kalman-based [19–21] approaches.

3.2 Bayesian estimation framework

In the sequel, we will often consider the processing of a single frame and therefore omit the frame index i. The speech enhancement problem can be formulated as a statistical estimation problem in which the clean speech, the noise and the noisy speech spectral coefficients $\underline{X}_k$, $\underline{W}_k$ and $\underline{Y}_k$ are represented as random variables[4]. In the Bayesian estimation approach [78] for speech enhancement, we wish to estimate the value of the clean speech spectrum $\underline{X}_k$ as a function of the noisy speech spectrum $\underline{Y}_k$. To do so, we define a distance metric, or cost function, between $\underline{X}_k$ and its estimator $\hat{\underline{X}}_k \equiv \hat{X}_k(\underline{Y}_k)$ and try to minimize the risk $\mathfrak{R}$ defined as the expectation of the chosen cost function [78]:

$$\mathfrak{R} \triangleq E\{C(\underline{X}_k, \hat{\underline{X}}_k)\} = \int\int C(X_k, \hat{X}_k) f_{\underline{X}_k, \underline{Y}_k}(X_k, Y_k) dX_k dY_k \tag{3.10}$$

where E denotes statistical expectation, $C(X_k, \hat{X}_k) \geq 0$ is the cost function and $f_{\underline{X}_k, \underline{Y}_k}(X_k, Y_k)$ is the joint probability density function (PDF) of $\underline{X}_k$ and $\underline{Y}_k$. The integral limits are not indicated in (3.10) but are with respect to the domains of the real and imaginary parts of X_k and Y_k which are from $-\infty$ to ∞. Furthermore, we can write $\mathfrak{R}$ as

$$\mathfrak{R} = \int f_{\underline{Y}_k}(Y_k) \int C(X_k, \hat{X}_k) f_{\underline{X}_k|\underline{Y}_k}(X_k|Y_k) dX_k dY_k \tag{3.11}$$

where

$$f_{\underline{X}_k|\underline{Y}_k}(X_k|Y_k) \triangleq \frac{f_{\underline{X}_k, \underline{Y}_k}(X_k, Y_k)}{f_{\underline{Y}_k}(Y_k)} = \frac{f_{\underline{Y}_k|\underline{X}_k}(Y_k|X_k) f_{\underline{X}_k}(X_k)}{f_{\underline{Y}_k}(Y_k)} \tag{3.12}$$

[4] In the remainder of this chapter, for clarity of presentation, random variables will be distinguished by an underline. However, following a common practice in the speech processing literature, this distinction will not be made in subsequent chapters.

is the conditional PDF of $\underline{X}_k$ given $\underline{Y}_k = Y_k$ and $f_{\underline{X}_k}(X_k)$ and $f_{\underline{Y}_k}(Y_k)$ are the marginal PDF of $\underline{X}_k$ and $\underline{Y}_k$ respectively. In the Bayesian formalism, one often refers to $f_{\underline{X}_k|\underline{Y}_k}(X_k|Y_k)$ as the *a posteriori* (i.e. after observing $\underline{Y}_k$) PDF and to $f_{\underline{X}_k}(X_k)$ as the *a priori* PDF. From (3.11), we have that:

$$\mathfrak{R} \geq \int f_{\underline{Y}_k}(Y_k) \min_{\hat{X}_k} \left\{ \int C(X_k, \hat{X}_k) f_{\underline{X}_k|\underline{Y}_k}(X_k|Y_k) dX_k \right\} dY_k \tag{3.13}$$

where the minimum is taken over all possible estimates $\hat{X}_k(Y_k)$. For a given Y_k, we therefore only need to minimize the inner expectation in (3.13) to obtain the desired estimator. The Bayesian estimate $\hat{X}_k$ is thus obtained as:

$$\hat{X}_k \quad = \quad \arg\min_{\hat{X}_k} \left\{ \int C(X_k, \hat{X}_k) f_{\underline{X}_k|\underline{Y}_k}(X_k|Y_k) dX_k \right\}. \tag{3.14}$$

Two elements need to be chosen in Bayesian estimation: firstly, the cost function $C(X_k, \hat{X}_k)$ to quantify the similarity between the clean speech and its estimate; and secondly, the statistical models used to characterize the various signals.

In the frequency domain framework, the complex STFT coefficients of the clean speech and noise at a given frequency k are generally modeled as statistically independent, identically distributed random variables with zero-mean and complex circular Gaussian distributions:

$$f_{\underline{X}_k}(X_k) = \frac{1}{\pi \sigma^2_{X,k}} e^{-|X_k|^2/\sigma^2_{X,k}} \tag{3.15}$$

$$f_{\underline{W}_k}(W_k) = \frac{1}{\pi \sigma^2_{W,k}} e^{-|W_k|^2/\sigma^2_{W,k}} \tag{3.16}$$

where

$$\sigma^2_{X,k} = E\{|\underline{X}_k|^2\} = E\{\underline{\mathcal{X}}^2_k\}; \qquad \sigma^2_{W,k} = E\{|\underline{W}_k|^2\} \tag{3.17}$$

denotes the corresponding variances respectively. Moreover, Fourier coefficients of the clean speech or noise taken at different frequencies are assumed to be independent, which in the Gaussian framework is equal to

$$E\{X_k X_l^*\} = E\{W_k W_l^*\} = 0 \tag{3.18}$$

for $k \neq l$.

The use of a Gaussian statistical model is motivated by the central limit theorem since each Fourier expansion coefficient can be seen as a weighted sum of random variables resulting from the observed samples [4]. Other distributions were also proposed for the real and imaginary parts of the STFT coefficients [31,39], the STSA coefficients [31,79] and the complex STFT coefficients [80,81]. While it has been proposed that the Fourier expansion coefficients of speech signals may not be Gaussian-distributed, those assumptions are usually motivated by long-term averages of the speech signal which may not be applicable to specific short-time utterances. Moreover, many estimators using a Gaussian distribution do not have an analytical counterpart when using other distributions [48]. Therefore, we will consider only Gaussian distributed complex STFT coefficients in this thesis.

Finally, we note that as a consequence of (3.15), it can be shown [38] that the amplitude and phase of $\underline{X}_k$, i.e. $\underline{\mathcal{X}}_k = |\underline{X}_k|$ and $\underline{\theta}_k = \angle \underline{X}_k$, are independent with Rayleigh and uniform distributions respectively, i.e.

$$f_{\underline{\mathcal{X}}_k, \underline{\theta}_k}(\mathcal{X}_k, \theta_k) = \frac{\mathcal{X}_k}{\pi \sigma_{X,k}^2} e^{-\mathcal{X}_k^2 / \sigma_{X,k}^2}. \tag{3.19}$$

3.3 Bayesian estimators of the STFT

A simple cost function is the magnitude squared error between the complex STFT coefficients X_k and $\hat{X}_k$:

$$C(X_k, \hat{X}_k) = |X_k - \hat{X}_k|^2. \tag{3.20}$$

This cost function applied in (3.14) leads to the minimum mean squared error (MMSE) estimate of $\underline{X}_k$ which is well known to be [38]:

$$\hat{X}_k = E\left\{\underline{X}_k | Y_k\right\}. \tag{3.21}$$

i.e. the conditional expectation of $\underline{X}_k$ given the observation $\underline{Y}_k = Y_k$. Under the Gaussian statistical models presented previously, the MMSE estimator of $\underline{\hat{X}_k}$ is the well-known Wiener estimator [48]:

$$\hat{X}_k^{\text{Wiener}} = \frac{\sigma_{X,k}^2}{\sigma_{X,k}^2 + \sigma_{W,k}^2} Y_k. \tag{3.22}$$

In (3.22), the gain applied to Y_k is real and positive and the phase of $\hat{X}_k^{\text{Wiener}}$ is therefore the phase of the noisy signal Y_k. The Wiener estimator (3.22) attenuates the frequency coefficients of the noisy speech with low signal-to-noise ratio, i.e. $\text{SNR}_k \triangleq \sigma_{X,k}^2/\sigma_{W,k}^2 \ll 1$, while frequency components with high SNR_k are essentially unchanged. The use of Wiener filter in speech enhancement generally introduces little speech distortion, however, as mentioned earlier, it produces much musical noise.

3.4 Bayesian estimators of the STSA

It has been shown that the spectral amplitude is perceptually more relevant than the phase in speech processing [40,82]. It seems therefore more appropriate to find the estimate of the

spectral *amplitude* rather than that of the complex spectrum. In this section, we present Bayesian estimators of the STSA instead of the STFT.

The STFT of the noisy speech, clean speech and noise can be decomposed into their amplitude and phase components as:

$$Y_k = |Y_k|\, e^{j\measuredangle Y_k}, \tag{3.23}$$

$$X_k = \mathcal{X}_k e^{j\theta_k}, \tag{3.24}$$

$$W_k = |W_k|\, e^{j\measuredangle W_k}. \tag{3.25}$$

where the notation $\mathcal{X}_k \triangleq |X_k|$ and $\theta_k \triangleq \measuredangle X_k$ is introduced for convenience. Since Y_k, X_k and W_k are STFT coefficients which are taken on finite windowed portion of the signal, the corresponding magnitude values $|Y_k|$, $\mathcal{X}_k$ and $|W_k|$ are commonly given the special terminology of short-time spectral amplitude (STSA) [3, 4]. Proceeding as in Section 3.2, the Bayesian estimator of the STSA, $\hat{\mathcal{X}}_k$, can be expressed as:

$$\hat{\mathcal{X}}_k^o \;=\; \arg\min_{\hat{\mathcal{X}}_k} \int_0^\infty C(\mathcal{X}_k, \hat{\mathcal{X}}_k) f_{\underline{\mathcal{X}}_k|\underline{Y}_k}(\mathcal{X}_k|Y_k) d\mathcal{X}_k. \tag{3.26}$$

This estimator will be combined with the phase of the noisy speech to obtain the estimate of the corresponding STFT coefficient:

$$\hat{X}_k^o = \hat{\mathcal{X}}_k^o e^{j\measuredangle Y_k}. \tag{3.27}$$

In [4], Ephraim and Malah studied the estimation of the complex exponential phase factor $e^{j\theta_k}$ of the clean speech. Using a constrained MMSE approach, they show that the optimal estimator of θ_k is the noisy phase of the clean speech, i.e. $\measuredangle Y_k$, itself. This justifies

the use of $\measuredangle Y_k$ in (3.27).

Let us now look at specific cost functions and the corresponding STSA estimators (3.26) obtained when considering the Gaussian statistical model discussed in Section 3.2 for the clean speech and noise.

3.4.1 MMSE STSA

Ephraim and Malah [4] proposed as cost function the squared error between the clean speech STSA and its estimate:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k - \hat{\mathcal{X}}_k)^2. \tag{3.28}$$

One could wonder what is the difference between the cost function (3.28) and the one given by (3.20). This is illustrated in Fig. 3.3, which shows the length of the difference between the STFT coefficients, i.e. $l_{X_k} \triangleq X_k - \hat{X}_k$, versus the length difference between the corresponding STSA, i.e. $l_{\mathcal{X}_k} = \mathcal{X}_k - \hat{\mathcal{X}}_k$. It can be observed that $l_{X_k} \geq l_{\mathcal{X}_k}$.

Using (3.28) in (3.26), it can be shown that:

$$\hat{\mathcal{X}}_k^o = E\{\underline{\mathcal{X}}_k | Y_k\} \tag{3.29}$$

$$= \int\int |X_k| f_{\underline{X}_k|\underline{Y}_k}(X_k|Y_k) dX_k \tag{3.30}$$

Using Bayes rule, this can be expanded as:

$$\hat{\mathcal{X}}_k^o = \frac{\int\int |X_k| f_{\underline{Y}_k|\underline{X}_k}(Y_k|X_k) f_{\underline{X}_k}(X_k) dX_k}{\int\int f_{\underline{Y}_k|\underline{X}_k}(Y_k|X_k) f_{\underline{X}_k}(X_k) dX_k} \tag{3.31}$$

This expression can be evaluated in closed form by considering the statistical model presented previously. Because the noise is additive (i.e. $\underline{Y}_k = \underline{X}_k + \underline{W}_k$) and independent

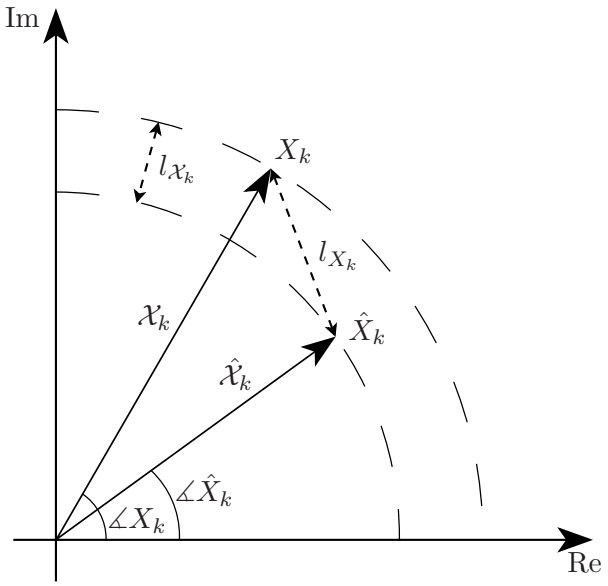

Fig. 3.3 Differences in length between STSA and STFT coefficients.

from the clean speech, it follows that:

$$f_{\underline{Y}_k|\underline{X}_k}(Y_k|X_k) = f_{\underline{W}_k}(Y_k - X_k) \tag{3.32}$$

and therefore

$$\hat{\mathcal{X}}_k^o = \frac{\int\int |X_k| f_{\underline{W}_k}(Y_k - X_k) f_{\underline{X}_k}(X_k) dX_k}{\int\int f_{\underline{W}_k}(Y_k - X_k) f_{\underline{X}_k}(X_k) dX_k} \tag{3.33}$$

Substituting the complex circular Gaussian PDF (3.15) and (3.16) into (3.33) and making a change of coordinates from rectangular to polar (i.e. $X_k = \mathcal{X}_k e^{j\theta_k}$) yields the MMSE STSA estimator [4]:

$$\hat{\mathcal{X}}_k^{\text{MMSE}} = G_k \, |Y_k| \tag{3.34}$$

$$G_k = \frac{\sqrt{\pi \upsilon_k}}{2\gamma_k} \exp\left(\frac{-\upsilon_k}{2}\right) \left[(1+\upsilon_k) \, I_0\left(\frac{\upsilon_k}{2}\right) + \upsilon_k I_1\left(\frac{\upsilon_k}{2}\right)\right] \tag{3.35}$$

where G_k is the gain applied to the spectral amplitudes of the noisy speech, $I_0(\cdot)$ and $I_1(\cdot)$

are the modified Bessel functions of zero and first order respectively [83] and

$$\upsilon_k = \frac{\xi_k}{1+\xi_k}\gamma_k, \quad \xi_k = \frac{\sigma^2_{X,k}}{\sigma^2_{W,k}}, \quad \gamma_k = \frac{|Y_k|^2}{\sigma^2_{W,k}}. \tag{3.36}$$

The parameters ξ_k and γ_k in (3.36) are respectively interpreted as the *a priori* and *a posteriori* Signal-to-Noise Ratio (SNR). The *a priori* SNR, ξ_k, acts as a long term estimator of the SNR whereas an *instantaneous* SNR can be defined as:

$$\gamma_k - 1 = \frac{|Y_k|^2 - \sigma^2_{W,k}}{\sigma^2_{W,k}}. \tag{3.37}$$

One can notice that the gain G_k (3.35) in the MMSE STSA estimator is a function of γ_k and ξ_k only. Therefore, only these two values need to be evaluated to compute the estimate. The estimation of these two parameters will be discussed in Section 3.5. Fig. 3.4 on p.45 shows the gain G_k of the MMSE STSA estimator (under the case $\beta = 1$)[5] as a function of the instantaneous SNR, $\gamma_k - 1$, for $\xi_k = 0$ dB. It is important to note that a smaller gain G_k will remove more background noise but will also introduce more speech distortion. Combined with the decision-directed approach to estimate ξ_k, which we will discuss later in Subsection 3.5.1, the MMSE STSA estimator has a residual noise that is much whiter than that of the Wiener estimator [4].

As observed from (3.35), the MMSE STSA estimator requires the computation of Bessel functions which can be updated by table look-up and/or interpolation techniques. However, some computationally efficient alternatives to the MMSE STSA estimator have been proposed in [84] that do not make use of Bessel functions. They are based on either maximum a posteriori (MAP) or MMSE estimation of the spectral power and lead to gains

[5]Parameter β will be explained in Subsection 3.4.3.

similar to the ones obtained with the MMSE STSA estimator while having much simpler expressions. The authors of [84] report an enhancement in speech quality similar to the one obtained by the MMSE STSA estimator but with a much reduced computational demand.

3.4.2 MMSE log-STSA (LSA)

Based on the assumption that the human auditory system performs a logarithmic compression of the STSA, and therefore that the logarithm of the STSA is more perceptually relevant than the STSA [85], the MMSE of the logarithm of the STSA was proposed in [28]. Its associated cost function is given by:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = [\log_z(\mathcal{X}_k) - \log_z(\hat{\mathcal{X}}_k)]^2 \tag{3.38}$$

where $\log_z(\cdot)$ is the logarithm to some base z. When (3.38) is used in (3.26), we obtain:

$$\hat{\mathcal{X}}_k^o = \exp[E\{\ln \underline{\mathcal{X}}_k | Y_k\}]. \tag{3.39}$$

which is independent of the base z chosen in (3.38) and where $\ln(\cdot)$ is the natural logarithm. Solving (3.39) with the previously mentioned statistical model leads to the MMSE log-STSA estimator, referred to as LSA in this thesis, for which the associated gain is given by:

$$G_k = \frac{\upsilon_k}{\gamma_k} \exp\left\{ \frac{1}{2} \int_{\upsilon_k}^{\infty} \frac{e^{-t}}{t} dt \right\}. \tag{3.40}$$

Fig. 3.4 also shows the gain G_k of the LSA estimator (under the case $\beta \to 0^+$) as a function of the instantaneous SNR, $\gamma_k - 1$, for $\xi_k = 0$ dB. This estimator results in a slightly higher speech distortion but lower residual noise than MMSE STSA. This can be explained

by the higher suppression (i.e. smaller gain) provided by the LSA estimator [28] as can be observed in Fig. 3.4. As for the MMSE STSA, the residual noise of the LSA estimator is white when combined with the decision-directed approach to estimate ξ_k.

which we will discuss later in Subsection 3.5.1, the MMSE STSA estimator has a residual noise that is much whiter than that of the Wiener estimator [4].

3.4.3 β-order STSA MMSE (β-SA)

The MMSE STSA estimator was generalized under the β-order STSA MMSE estimator in [29], denoted by β-SA in this thesis[6]. The β-SA cost function is given by:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k^\beta - \hat{\mathcal{X}}_k^\beta)^2 \tag{3.41}$$

where the exponent β is a real parameter whose purpose is to control the associated estimator gain function and, consequently, the trade-off between speech distortion and noise reduction. Only the case $\beta > 0$ was considered in [29].

With (3.41) in (3.26), we have:

$$\hat{\mathcal{X}}_k^o = \sqrt[\beta]{E\{\mathcal{X}_k^\beta | Y_k\}}. \tag{3.42}$$

for which the gain of the corresponding β-SA estimator is expressible as:

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \left[\Gamma\left(\frac{\beta}{2} + 1\right) M\left(-\frac{\beta}{2}, 1; -\upsilon_k\right) \right]^{1/\beta} \tag{3.43}$$

[6] An equivalent estimator for the power spectra of the clean speech, $\hat{\mathcal{X}}_k^2$, was also derived in [30] and termed *Generalized MMSE*.

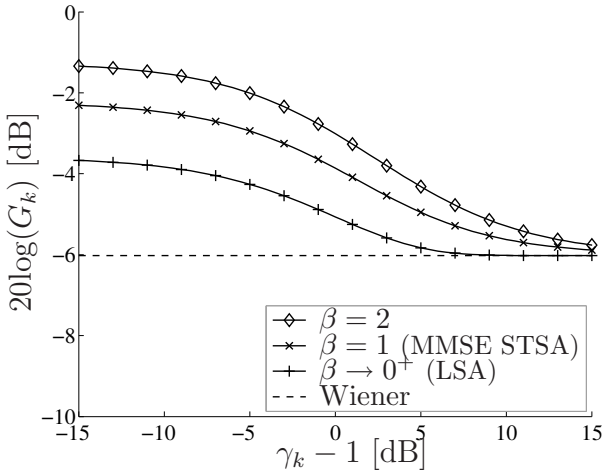

Fig. 3.4 β-SA estimator gain ($20\log G_k$) versus instantaneous SNR, $\gamma_k - 1$, for several β values ($\xi_k = 0$ dB).

In this expression, $\Gamma(h)$ is the gamma function [83]

$$\Gamma(h) = \int_0^\infty t^{h-1} e^{-t} dt \tag{3.44}$$

and $M(a, b; z)$ is the confluent hypergeometric function [83]

$$M(a, b; z) = 1 + \frac{a}{b}\frac{z}{1!} + \frac{a(a+1)}{b(b+1)}\frac{z^2}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)}\frac{z^3}{3!} + ... \tag{3.45}$$

As can be readily observed from (3.41), the β-SA estimator is equivalent to the MMSE STSA estimator when $\beta = 1$. Furthermore, it was observed in [29] based on gain curve analysis, that when $\beta \to 0^+$, the β-SA estimator corresponds to the LSA estimator. Therefore, the MMSE STSA and LSA estimators appear to be subsets of the more general β-SA estimator.

Fig. 3.4 shows the gain G_k as a function of the instantaneous SNR, $\gamma_k - 1$, for the β-SA estimator with several values of β and for the Wiener estimator (3.22) with $\xi_k = 0$ dB. As can be observed, the β-SA estimator tends to the Wiener estimator when the instantaneous

SNR is large. Moreover, when β decreases towards 0, the gain decreases and therefore the estimator inevitably removes more noise but at the same time produces also more speech distortion. You *et al.* [29] proposed adapting the value of β according to each frame's SNR. They assigned smaller values of β to frames having smaller SNRs, therefore reducing more noise, and larger values of β to frames having larger SNRs, therefore limiting the speech distortion for those frames. They argued that their method outperforms many existing estimators including the MMSE STSA and LSA estimators.

Moreover, You *et al.* [86] have tried to render the cost function (3.41) more perceptually significant by exploiting the masking properties of the ear. In fact, they proposed to modify the values of β in the β-SA estimator according to both the values of the masking threshold for each frequency and the frame SNR as given by the following empirical function:

$$\beta_{k,i} = \tau_0 + \tau_1 \mathrm{SNR}_i + \tau_2 A_{k,i} + \tau_3 \mathrm{SNR}_i A_{k,i} \tag{3.46}$$

where as before, k is the frequency index, i is the frame index, SNR_i is the SNR for frame i, $A_{k,i}$ takes account for the masking threshold and τ_a, $a \in \{0, 1, 2, 3\}$, are empirically determined coefficients. On the one hand, if the masking threshold is high, a large value of β is chosen to limit the amount of speech distortion introduced; on the other hand, if the masking threshold is low, they choose a low value of β to further reduce the noise. As previously, they also choose a larger value of β for a frame with high SNR and vice versa. In addition, the β values in [86] are constrained between 0.001 and 4. The resulting estimator is found to surpass the LSA estimator.

3.4.4 Weighted euclidian (WE)

Some speech distance measures, such as the well known Itakura-Saito measure, have been known for a long time to be more perceptually relevant than others [87]. Loizou [6] studied the use of several perceptually meaningful distance measures such as the weighted likelihood ratio, the Itakura-Saito distance measure and one of its variants, the COSH distance measure, which were used as cost functions in a Bayesian estimation setting. Additionally he proposed a cost function that is based on the perceptually-weighted error criterion used in speech coding which he termed weighted Euclidian (WE). We describe the WE estimator here and the COSH in the next subsection.

Let us first look at the WE estimator whose cost function has the following form:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \mathcal{X}_k^p \left(\mathcal{X}_k - \hat{\mathcal{X}}_k\right)^2 \tag{3.47}$$

where p is a real parameter. This cost function corresponds to the one of the MMSE STSA estimator when $p = 0$. When (3.47) is used in (3.26), the latter reduces to:

$$\hat{\mathcal{X}}_k^o = \frac{E\{\mathcal{X}_k^{p+1}|Y_k\}}{E\{\mathcal{X}_k^p|Y_k\}} \tag{3.48}$$

This expression can be evaluated in closed-form and the corresponding WE estimator has the following gain:

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \frac{\Gamma\left(\frac{p+1}{2}+1\right)}{\Gamma\left(\frac{p}{2}+1\right)} \frac{M\left(-\frac{p+1}{2}, 1; -\upsilon_k\right)}{M\left(-\frac{p}{2}, 1; -\upsilon_k\right)}. \tag{3.49}$$

which is valid for $p > -2$.

The WE estimator takes advantage of the masking properties of the ear. In fact, for $p < 0$, the cost function in (3.47) forces a better clean speech estimation in regions where the STSA is smaller, and therefore less likely to mask noise remaining in the clean speech

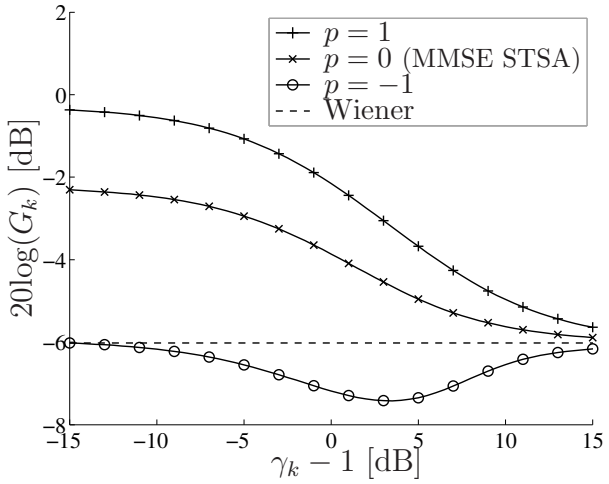

Fig. 3.5 WE estimator gain ($20\log G_k$) versus instantaneous SNR, $\gamma_k - 1$, for several p values ($\xi_k = 0$ dB).

estimate. Similarly to β in the β-SA estimator, p was also found to control the trade-off between speech distortion and noise reduction [6].

Fig. 3.5 plots the gain G_k as a function of the instantaneous SNR, $\gamma_k - 1$, for the WE estimator with several values of p and for the Wiener estimator ($\xi_k = 0$ dB). It can be observed that a smaller value of p will produce a smaller gain and therefore more noise reduction but also greater speech distortion. The value of $p = -1$ has been suggested in [6] as a good compromise between the desired noise reduction and the speech distortion introduced. It can also be seen in Fig. 3.5 that, as for the β-SA, the WE estimator tends to the Wiener estimator for large instantaneous SNR; this was formally proven in [6].

3.4.5 COSH and weighted COSH (WCOSH)

The COSH distance measure is a variant of the well-known Itakura-Saito distortion measure which, as opposed to the latter, is symmetric in the sense that $C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = C(\hat{\mathcal{X}}_k, \mathcal{X}_k)$. It was also proposed in [6] as a cost function to be used in a speech enhancement Bayesian

estimator. Its associated cost function is:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \frac{1}{2}\left(\frac{\mathcal{X}_k}{\hat{\mathcal{X}}_k} + \frac{\hat{\mathcal{X}}_k}{\mathcal{X}_k}\right) - 1. \tag{3.50}$$

A generalization of the COSH cost function, the weighted COSH (WCOSH), was also proposed in [6] for which the cost function is:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \left(\frac{\mathcal{X}_k}{\hat{\mathcal{X}}_k} + \frac{\hat{\mathcal{X}}_k}{\mathcal{X}_k} - 1\right)\mathcal{X}_k^q \tag{3.51}$$

where q is a real parameter. The associated WCOSH estimator is obtained when using (3.51) in (3.26) and is given by:

$$\hat{\mathcal{X}}_k^o = \sqrt{\frac{E\{\hat{\underline{\mathcal{X}}}_k^{q+1}|Y_k\}}{E\{\hat{\underline{\mathcal{X}}}_k^{q-1}|Y_k\}}}. \tag{3.52}$$

with the gain of the corresponding estimator being [6]:

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k}\sqrt{\frac{\Gamma\left(\frac{q+3}{2}\right) M\left(-\frac{q+1}{2}, 1, -\upsilon_k\right)}{\Gamma\left(\frac{q+1}{2}\right) M\left(-\frac{q-1}{2}, 1, -\upsilon_k\right)}} \tag{3.53}$$

which is valid for $q > -1$. In the case $q = 0$, the resulting estimator is identical to the COSH estimator.

Fig. 3.6 shows the gain G_k of the WCOSH estimator as a function of the instantaneous SNR, $\gamma_k - 1$, for several values of q and for $\xi_k = 0$ dB. Again, we can see that q will control the trade-off between speech distortion and noise reduction since a smaller q will produce a smaller gain G_k. As the β-SA and WE estimators, the WCOSH estimator also tends to Wiener's for large instantaneous SNR.

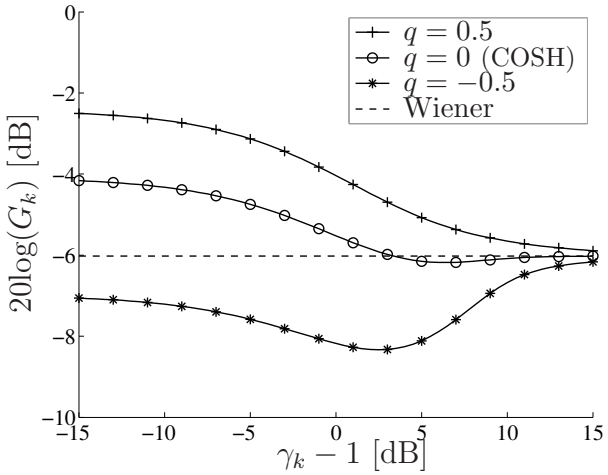

Fig. 3.6 WCOSH estimator gain ($20\log G_k$) versus instantaneous SNR, $\gamma_k - 1$, for several q values ($\xi_k = 0$ dB).

One major conclusion of the study in [6] was that the estimators which over-emphasized the spectral peaks in the cost function, such as the MMSE STSA or WE with $p > 0$, performed the worst. According to [6], this is due to the fact that those estimators produced a small estimation error at spectral peaks where the noise is more likely masked while they produced larger estimation errors in the spectral valleys. It was also noticed in [6] that the estimators that emphasized spectral valleys, such as the WE with $p < 0$, performed the best. In fact, these estimators implicitly exploit the auditory masking properties of the ear since in this case, the large estimation errors near the spectral peaks are properly masked. Among all studied estimators in [6], only the WE with a proposed value of $p = -1$, which therefore emphasized the spectral valleys, outperformed the LSA estimator.

3.4.6 Summary of Bayesian STSA estimators

Table 3.1 summarizes the different Bayesian STSA cost functions discussed in this section along with their corresponding gains G_k, the MMSE STSA gain G_k is expressed in Table 3.1 in a form equivalent to (3.35) where the Bessel functions are replaced by gamma and

Table 3.1 Cost functions with corresponding gains G_k for several existing Bayesian STSA estimators.

	$C(\mathcal{X}_k, \hat{\mathcal{X}}_k)$	G_k
MMSE STSA [4]	$(\mathcal{X}_k - \hat{\mathcal{X}}_k)^2$	$\frac{\sqrt{\upsilon_k}}{\gamma_k}\Gamma(1.5)M\left(-0.5, 1; -\upsilon_k\right)$
LSA [28]	$(\log \mathcal{X}_k - \log \hat{\mathcal{X}}_k)^2$	$\frac{\upsilon_k}{\gamma_k} \exp\left\{\frac{1}{2}\int_{\upsilon_k}^{\infty} \frac{e^{-t}}{t} dt\right\}$
COSH [6]	$\frac{1}{2}\left(\frac{\mathcal{X}_k}{\hat{\mathcal{X}}_k} + \frac{\hat{\mathcal{X}}_k}{\mathcal{X}_k}\right) - 1$	$\frac{\sqrt{\upsilon_k}}{\gamma_k}\sqrt{\frac{1}{2}\frac{M(-0.5,1;-\upsilon_k)}{M(0.5,1;-\upsilon_k)}}$
β-SA [29]	$(\mathcal{X}_k^{\beta} - \hat{\mathcal{X}}_k^{\beta})^2$	$\frac{\sqrt{\upsilon_k}}{\gamma_k}\left[\Gamma(\frac{\beta}{2}+1)M\left(-\frac{\beta}{2}, 1; -\upsilon_k\right)\right]^{1/\beta}$
WE [6]	$\mathcal{X}_k^p(\mathcal{X}_k - \hat{\mathcal{X}}_k)^2$	$\frac{\sqrt{\upsilon_k}}{\gamma_k}\frac{\Gamma\left(\frac{p+1}{2}+1\right)}{\Gamma\left(\frac{p}{2}+1\right)}\frac{M\left(-\frac{p+1}{2},1;-\upsilon_k\right)}{M\left(-\frac{p}{2},1;-\upsilon_k\right)}$
WCOSH [6]	$\left(\frac{\mathcal{X}_k}{\hat{\mathcal{X}}_k} + \frac{\hat{\mathcal{X}}_k}{\mathcal{X}_k} - 1\right)\mathcal{X}_k^q$	$\frac{\sqrt{\upsilon_k}}{\gamma_k}\sqrt{\frac{\Gamma\left(\frac{q+3}{2}\right)M\left(-\frac{q+1}{2},1;-\upsilon_k\right)}{\Gamma\left(\frac{q+1}{2}\right)M\left(-\frac{q-1}{2},1;-\upsilon_k\right)}}$

confluent hypergeometric functions.

3.5 Parameter estimation

The Bayesian estimators of the STSA presented in the previous section are function of two common parameters: the *a priori* SNR ξ_k and the *a posteriori* SNR γ_k. In Subsection 3.5.1, we present methods to estimate ξ_k, while in Subsection 3.5.2, we present methods for the estimation of γ_k.

3.5.1 *A priori* SNR estimation

We note from (3.36) and (3.37) that ξ_k can be obtained by either $\xi_k = \sigma_{X,k}^2/\sigma_{W,k}^2$ or $\xi_k = E\{\gamma_k - 1\}$. Accordingly, Ephraim and Malah [4] proposed a *decision-directed* approach

to estimate ξ_k that combines these two alternative expressions:

$$\hat{\xi}_{k,i} = \tau \frac{\hat{\mathcal{X}}^2_{k,i-1}}{\sigma^2_{W,k,i-1}} + (1-\tau)\max\left[\gamma_{k,i} - 1, 0\right] \tag{3.54}$$

where $\hat{\mathcal{X}}_{k,i-1}$ represents the STSA estimate at frequency k from a previous frame, $\sigma^2_{W,k,i-1}$ is the variance of the noise at frequency k from a previous frame, $\gamma_{k,i}$ is the *a posteriori* SNR for the current frame and τ is a weighting parameter with typical values in the range $0.95 \leq \tau < 1$. The $\max[\cdot,\cdot]$ operator is introduced to remove the possibility of negative values in the instantaneous SNR, i.e. $\gamma_{k,i} - 1$. This nonlinear smoothing procedure has the great advantage of eliminating large variations across successive frames and therefore reducing the musical noise [88]. However, it will respond slowly to an abrupt increase in the instantaneous SNR. In fact, for the large values of τ typically used, $\hat{\xi}_{k,i}$ is not well estimated during the onset of speech. If the actual SNR varies quickly, $\hat{\xi}_{k,i}$ will not be able to adapt accordingly.

The decision-directed approach depends greatly on the STSA estimation of the previous frame $\hat{\mathcal{X}}_{k,i-1}$. The resulting bias, towards the STSA of the previous frame instead of that of the current one, produces an annoying reverberation artifact in the STSA estimation which is especially prominent when the overlap between frames is of 50% instead of the 75% overlap used in [4]. In order to overcome this problem, Plapous *et al.* in [89] proposed a two-step noise reduction approach were a gain function is first evaluated using the decision-directed approach, this gain is then used to compute an estimation of the power spectral density of the speech $E\{\mathcal{X}^2_{k,i}\}$ that is used to compute the *a priori* SNR ξ_k.

Other improvements to the decision-directed approach were also proposed in [27] and [74]. In [27], the parameter τ in (3.54) is optimally derived by minimizing $E\{(\hat{\xi}_{k,i} - \underline{\xi}_{k,i})^2 | \tilde{\xi}_{k,i-1}\}$ where $\tilde{\xi}_{k,i-1} = \hat{\mathcal{X}}^2_{k,i-1}/\sigma^2_{W,k,i-1}$. The authors report better speech enhance-

ment results than when using the original decision-directed approach. In [74], a noncausal estimator of the *a priori* SNR is proposed. This estimator has access to subsequent noisy speech samples. This added knowledge gives the estimator the advantage of being able to better discriminate between speech onsets and noise irregularities than the decision-directed approach in [4]. It demands however that the application can tolerate a certain amount of delay.

3.5.2 *A posteriori* SNR and noise variance estimation

As given by (3.36), the *a posteriori* SNR is equal to $\gamma_k = |Y_k|^2/\sigma^2_{W,k}$. Since $|Y_k|^2$ is the known observation, only $\sigma^2_{W,k}$ needs to be estimated. The noise variance estimate is particularly important in most single-channel speech enhancement approaches and not only in the Bayesian approach.

If the noise is stationary, its statistics can simply be evaluated from any speech free frame. The most straightforward method to evaluate the noise statistics is to identify a time frame when only noise is present and estimate the noise statistics from that frame. In order to identify frames where there is only noise, a voice activity detector (VAD) can be used [90–93]. VAD's are present in many speech codecs such as the EVRC [62].A VAD detects the presence of speech in a noisy speech signal. One common assumption in VAD's is that the energy of the input signal will be higher when there is speech than when there is only background noise. The energy can therefore be estimated and a threshold is set over which it is decided that speech is present. VAD's are therefore less accurate when the SNR is low.

Alternatively, for non-stationary noises, the noise statistics need to be more frequently updated. In that regard, many algorithms have also been proposed that use a soft-decision approach where the noise is updated continually, i.e. whether speech is present or not,

therefore not needing a VAD to detect speech-free periods. Some soft-decision approaches [94, 95] are based on the observation that the power spectral density estimate, even during speech presence, frequently drops to values that are representative of the noise power level. The associated algorithms therefore assume that the speech energy is close to or equal to zero during speech pauses or in between words or syllables. Furthermore, since the minimum value considered here will necessarily be smaller than the mean of the noise power, the estimator is biased. A compensation scheme is therefore introduced in the estimator to obtain an unbiased noise estimator.

In [96], the noise estimate is obtained by averaging past spectral power values and using a smoothing parameter that is adjusted by the signal presence probability in subbands. The presence of speech in subbands is determined by a minimum energy scheme similar to the one in [94, 95]. This method is further improved in [97] where a two iteration procedure is adopted. In the first iteration, a rough voice activity detection is performed in each frequency band, then, smoothing in the second iteration is used to exclude relatively strong speech components, making the minimum tracking during speech activity robust.

According to [98], both the methods of [95] and [97] use a finite temporal window that can be quite long. This introduces some delay in the tracking of the noise statistics and has the effect of introducing a slow response to rapidly increasing noise level under non-stationary noise conditions. The authors of [98] propose an approach that does not suffer from this drawback and where the noise variance estimate, $\hat{\sigma}^2_{W,k,i}$, is updated recursively with the MMSE estimate of the current noise power:

$$\hat{\sigma}^2_{W,k,i} = \tau_w \sigma^2_{W,k-1,i} + (1 - \tau_w) D_{i,k} \tag{3.55}$$

where $D_{i,k}$ is the MMSE estimate of $|\underline{W}_{k,i}|^2$. They claim to adequately track fast changes

in noise power levels up to 10 dB/s.

3.6 Summary

In this chapter, we presented an overview of the Bayesian approach for speech enhancement with an emphasis on Bayesian STSA estimators in the frequency domain. In particular, we exposed several Bayesian estimators of the STSA including the MMSE STSA, LSA, β-SA, WE, COSH and WCOSH estimators. In order to gain a better understanding of the properties of the class of Bayesian STSA estimators and, more importantly, to improve their performance, new estimators will be presented in the remainder of this thesis that extend and build upon those reviewed in this chapter.

Chapter 4

Further analysis and extension of the β-SA estimator

While providing insight into the operation of the β-SA estimator, the analysis in [29] only considered the case $\beta > 0$ and relied on empirical observations in establishing a link between the β-SA and LSA estimators. In this chapter, we extend the scope of the analysis in [29] to address the above limitations.

In Section 4.1, we briefly recapitulate the underlying assumption and cost function of the β-SA estimator for convenience to the reader. In Section 4.2, we first show that the expression obtained for the β-SA estimator remains, in fact, valid for $\beta > -2$. We then provide an interpretation and analysis of the estimator for β values in the range $-2 < \beta < 0$. In Section 4.3, we provide an original formal mathematical proof of the equivalence between the special case $\beta \to 0^+$ and the LSA estimator.

4.1 Problem formulation

We start by reviewing briefly some elements of the β-SA estimator introduced in Section 3.4.3. In this chapter, and also in Chapters 5 and 6, we assume the additive noise model in the STFT domain introduced in Section 3.1.2, which can be formulated for a given frame as:

$$Y_k = X_k + W_k \tag{4.1}$$

where Y_k, X_k and W_k denote the STFT coefficients of the noisy speech, clean speech and noise respectively. As in Section 3.4, we wish to find the estimator $\hat{X}_k^o = \hat{\mathcal{X}}_k^o e^{j\measuredangle Y_k}$ where $\hat{\mathcal{X}}_k^o$ is the estimator of the STSA coefficient of X_k and $\measuredangle Y_k$ is the phase of Y_k. In the Bayesian formalism, $\hat{\mathcal{X}}_k^o$ is obtained as the minimum solution to $E\{C(\mathcal{X}_k, \hat{\mathcal{X}}_k)\}$. The β-SA cost function $C(\mathcal{X}_k, \hat{\mathcal{X}}_k)$, defined in (3.41), is given by the following:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k^\beta - \hat{\mathcal{X}}_k^\beta)^2 \tag{4.2}$$

where the exponent β is a real parameter whose purpose is to control the associated estimator gain function and, consequently, the trade-off between speech distortion and noise reduction. For example, a value of β close to 1 will produce a gain G_k closer to 1 and therefore less noise reduction and speech distortion than a value close to 0 which will produce a smaller gain.

The β-SA estimator was originally proposed in [29] where it was analyzed for values of $\beta > 0$. However, it can be shown that the expression obtained therein for the β-SA estimator remains valid for $\beta > -2$. These values of β may reveal some advantages in terms of the quality of the corresponding enhanced speech over their positive β value counterparts. Moreover, the special case $\beta \to 0^+$ was empirically shown to correspond to

the LSA estimator (Subsection 3.4.2) in [29] through the comparison of gain curves. A formal mathematical proof of that equivalence is missing.

4.2 The case $\beta < 0$

Let us start by showing that the expression for the β-SA gain given by (3.43) remains valid for $\beta > -2$. To obtain the β-SA estimator, one needs to evaluate (3.42), repeated here for convenience:

$$\hat{\mathcal{X}}_k^o = \sqrt[\beta]{E\{\mathcal{X}_k^\beta | Y_k\}}. \tag{4.3}$$

Considering the statistical model described in Section 3.2, in which $\mathcal{X}_k$ obeys a Rayleigh distribution, a general expression for the expectation appearing in (4.3) is obtained in [29] as:

$$E\{\mathcal{X}_k^\beta | Y_k\} = \frac{\int_0^\infty \mathcal{X}_k^{\beta+1} \exp\{-\mathcal{X}_k^2/\varsigma_k\} I_o(2\mathcal{X}_k\sqrt{\upsilon_k/\varsigma_k}) d\mathcal{X}_k}{\int_0^\infty \mathcal{X}_k \exp\{-\mathcal{X}_k^2/\varsigma_k\} I_o(2\mathcal{X}_k\sqrt{\upsilon_k/\varsigma_k}) d\mathcal{X}_k} \tag{4.4}$$

where $I_o(\cdot)$ is the modified Bessel function of order zero, υ_k is defined in (3.36) and $\varsigma_k = (1/\sigma_{W,k}^2 + 1/\sigma_{X,k}^2)^{-1}$. To solve (4.4) requires the following relation as given by (6.631.1) in [83]:

$$\int_0^\infty x^a e^{-bx^2} J_c(mx) dx = \frac{m^c \Gamma(a/2 + c/2 + 1/2)}{2^{c+1} b^{(a+c+1)/2} \Gamma(c+1)} M\left(\frac{a+c+1}{2}, c+1; -\frac{m^2}{4b}\right) \tag{4.5}$$

where $m > 0$, $a \in \mathbb{C}$, $b \in \mathbb{C}$, $c \in \mathbb{C}$ and $J_c(mx)$ is a Bessel function of the first kind with $J_c(jx) = j^c I_c(x)$. Eq. (4.5) is valid for $\Re(b) > 0$ and $\Re(a+c) > -1$ where $\Re(x)$ indicates the real part of x.

It is shown in [29] that using (4.5) in both the numerator and denominator of (4.4) with

appropriate parameter values, one can solve (4.4) and therefore obtain the β-SA estimator:

$$\hat{\mathcal{X}}_k^o = G_k|Y_k| \tag{4.6}$$

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k}\left[\Gamma\left(\frac{\beta}{2}+1\right)M\left(-\frac{\beta}{2},1;-\upsilon_k\right)\right]^{1/\beta} \tag{4.7}$$

where $\Gamma(\cdot)$ is the gamma function defined in (3.44), $M(u,t;z)$ is the confluent hypergeometric function defined in (3.45) and γ_k is defined in (3.36). In particular, to evaluate the integral in the numerator of (4.4) requires the following parameter values in (4.5): $a=\beta+1$, $b=1/\varsigma_k$ and $c=0$. This implies that (4.7) is valid firstly for $\Re(1/\varsigma_k)>0$, which is always true, and secondly for $\Re(\beta+1)>-1$ or equivalently $\beta>-2$. However, only the range $\beta>0$ was considered in [29]. In fact, there is no mention of possible negative β values and, furthermore, the analysis and evaluation of the β-SA estimator was only done for values of $\beta>0$. Therefore, the study of the β-SA estimator for the case $-2<\beta\leq 0$ remains an open issue and will be the subject of the remainder of this chapter.

4.2.1 A normalization interpretation

For negative β values, we have that $\beta=-|\beta|$ and the β-SA cost function (4.2), explicitly expressed as a function of β, becomes:

$$\begin{aligned} C(\mathcal{X}_k,\hat{\mathcal{X}}_k;\beta) &= \left(\frac{1}{\mathcal{X}_k^{|\beta|}}-\frac{1}{\hat{\mathcal{X}}_k^{|\beta|}}\right)^2 \\ &= \left(\frac{\hat{\mathcal{X}}_k^{|\beta|}-\mathcal{X}_k^{|\beta|}}{\mathcal{X}_k^{|\beta|}\hat{\mathcal{X}}_k^{|\beta|}}\right)^2 \\ &= \frac{C(\mathcal{X}_k,\hat{\mathcal{X}}_k;|\beta|)}{\left(\mathcal{X}_k\hat{\mathcal{X}}_k\right)^{2|\beta|}} \end{aligned} \tag{4.8}$$

From (4.8), we observe that using a negative value of β amounts to normalizing the cost function for positive β, i.e. $C(\mathcal{X}_k, \hat{\mathcal{X}}_k; |\beta|)$, by $\left(\mathcal{X}_k \hat{\mathcal{X}}_k\right)^{2|\beta|}$.

The denominator in (4.8) can be thought of as an approximation of the power spectrum of the desired speech, $E\{\mathcal{X}_k^2\}$, to which is applied an exponent $2\,|\beta|$. This normalization thus penalizes the estimation error more heavily when the power spectrum is small, which corresponds to spectral valleys, than when it is large, i.e. corresponding to spectral peaks. More noise will likely be audible in the speech spectral valleys than in the speech spectral peaks where it will more likely be masked by the speech. The β-SA estimator with $\beta < 0$ can therefore take advantage of the masking properties of the human ear by favoring a more accurate estimation of the speech in the spectral valleys. This behavior is thus similar to the one of the WE estimator with $p < 0$ (see Subsection 3.4.4).

4.2.2 Analysis of the β-SA estimator with $\beta < 0$

In this subsection, we first analyze the behavior of the gain in the β-SA estimator for $\beta < 0$ and then proceed with a study of the noise reduction and speech distortion introduced by the corresponding estimator.

Gain versus instantaneous SNR

Fig. 4.1 (a) and (b) shows numerical plot of the β-SA gain (4.7) versus the instantaneous SNR, $\gamma_k - 1$, for several values of β ($\beta \to 0$, $\beta = -0.5, -1$ and -1.5) and for $\xi_k = 0$ dB and $\xi_k = 10$ dB respectively. In connection with Fig. 3.4 for the case $\beta > 0$, we already observed that the gain G_k always decreases as β decreases. From Fig. 4.1, we note that this trend continues for negative values of β as well. However, while for $\beta > 0$ the gain is a monotonically decreasing function of $\gamma_k - 1$, it is not so anymore for $\beta < 0$. Furthermore, it was noted in [29] that the β-SA gain G_k exceeded but converged to the Wiener filter gain

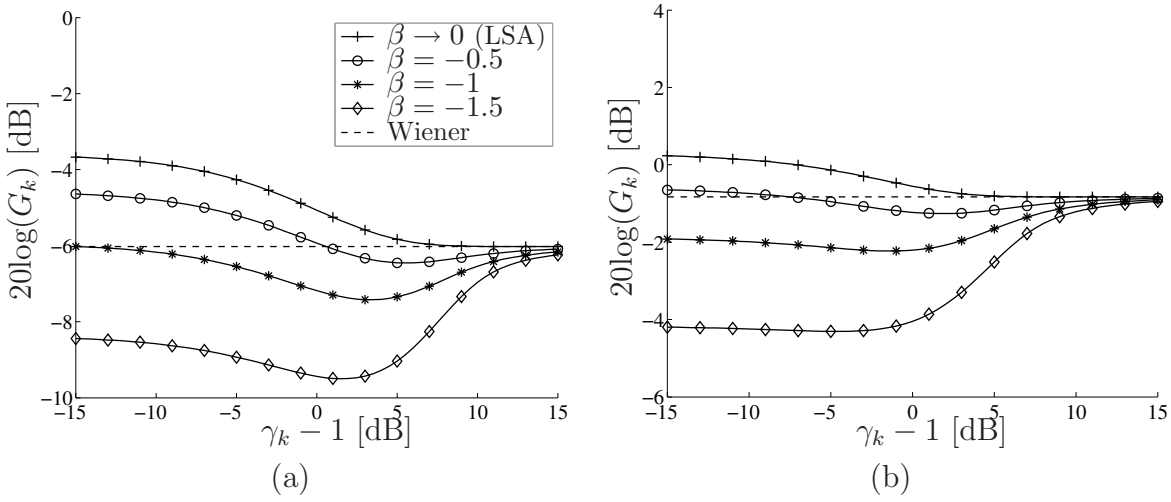

Fig. 4.1 β-SA estimator gain ($20\log G_k$) versus instantaneous SNR, $\gamma_k - 1$, for several values of $\beta < 0$ (a) $\xi_k = 0$ dB and (b) $\xi_k = 10$ dB.

as the instantaneous SNR increased. We see that it is still the case for $\beta < 0$; however, the β-SA gain can now become less than the Wiener filter's gain.

Noise reduction versus speech distortion

As observed in Fig. 4.1, the gain of the β-SA estimator decreases as β decreases. Therefore, for smaller β values, more noise reduction, but also more speech distortion should be expected. In order to study the speech distortion and noise reduction properties of the β-SA estimator over the extended range $\beta > -2$, we use the following speech distortion metric, $\eta_{\mathrm{SD}}(G_k)$, and noise reduction metric, $\eta_{\mathrm{NR}}(G_k)$, in the frequency domain[1]:

$$\eta_{\mathrm{SD}}(G_k) = \frac{E\left\{[\mathcal{X}_k - G_k\mathcal{X}_k]^2\right\}}{E\{\mathcal{X}_k^2\}} \tag{4.9}$$

$$\eta_{\mathrm{NR}}(G_k) = \frac{E\{|W_k|^2\}}{E\left\{[G_k\,|W_k|]^2\right\}} \tag{4.10}$$

[1]These metrics are adapted from their time domain counterparts in [70].

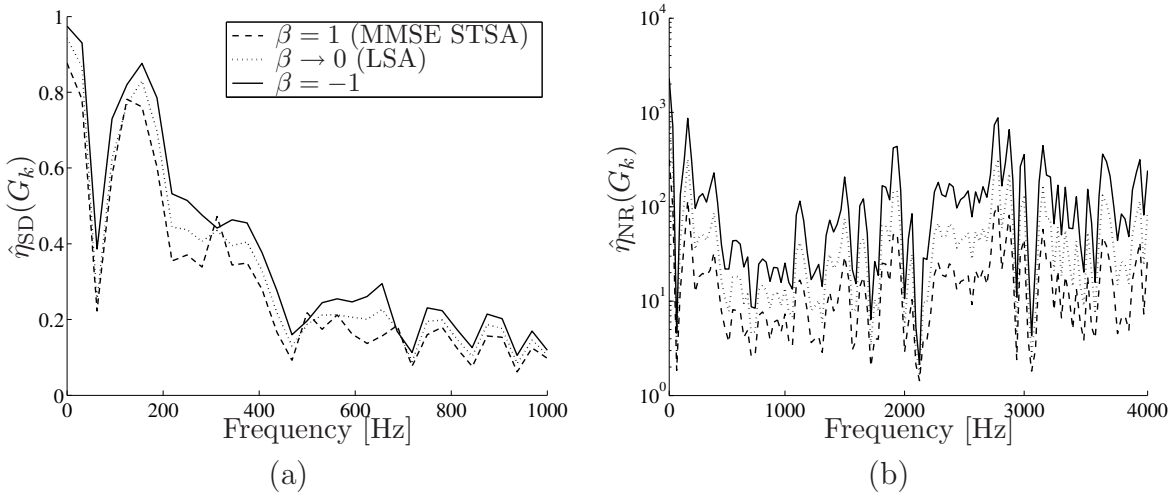

Fig. 4.2 (a) Speech distortion $\hat{\eta}_{\mathrm{SD}}(G_k)$ vs. frequency (0 - 1000 Hz) (b) Noise reduction $\hat{\eta}_{\mathrm{NR}}(G_k)$ vs. frequency (0 - 4000 Hz) (white noise, SNR = 0 dB).

where the clean speech and noise variances, $\sigma^2_{X,k}$ and $\sigma^2_{W,k}$ respectively, are given by (3.17). In (4.9), $\eta_{\mathrm{SD}}(G_k)$ measures the normalized clean speech distortion energy and, therefore, its value increases for increasing speech distortions. In (4.10), $\eta_{\mathrm{NR}}(G_k)$ computes the ratio of the original noise power to the power of the residual noise in the enhanced speech, its value increases for increasing noise reduction.

Fig. 4.2 plots estimations of $\eta_{\mathrm{SD}}(G_k)$ and $\eta_{\mathrm{NR}}(G_k)$, respectively $\hat{\eta}_{\mathrm{SD}}(G_k)$ and $\hat{\eta}_{\mathrm{NR}}(G_k)$, versus the frequency in Hz for several β values (white noise, SNR = 0 dB). These estimations were obtained by, firstly, computing the β-SA gain for each frame of 30 different sentences and, secondly, performing an average of the arguments of the expectations in (4.9) and (4.10) over all frames of the 30 sentences. As expected, we observe that both the speech distortion and noise reduction increase as β is decreased from $\beta = 1$ (corresponding to the MMSE STSA estimator) to the extended case $\beta = -1$. The case $\beta \to 0$ (corresponding to the LSA estimator) exhibits an intermediate behavior.

The use of negative β thus enables an extension of the trade-off between speech distor-

tion and noise reduction as compared with the strictly positive β case.

4.3 The limiting case $\beta \to 0$

In [29] it is argued that the β-SA estimator with $\beta \to 0^+$ is equivalent to the LSA estimator (see Subsection 3.4.2). However, this is based solely on the empirical comparison of gain curves versus instantaneous SNR obtained from the LSA estimator on the one hand and from the β-SA estimator with a value of β close to zero on the other hand. In this section, we provide an original mathematical proof that the β-SA estimator with $\beta \to 0$ is indeed equivalent to the LSA estimator.

We begin by expressing the β-SA estimator gain (4.7) in the form:

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \exp\left\{\frac{1}{\beta} \ln\left[\Gamma\left(\frac{\beta}{2}+1\right) M\left(-\frac{\beta}{2}, 1; -\upsilon_k\right)\right]\right\}. \tag{4.11}$$

Using (8.342.1) from [83]:

$$\ln\Gamma(z+1) = -\gamma z + \sum_{l=2}^{\infty} (-1)^l \frac{z^l}{l} \zeta(l), \quad |z| < 1 \tag{4.12}$$

where γ is Euler's constant and $\zeta(l)$ is Weierstrass's zeta function (see (8.17) in [83]), we have that:

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \exp\left\{-\frac{\gamma}{2} + \frac{1}{\beta} \sum_{l=2}^{\infty} \frac{1}{l} \left(\frac{-\beta}{2}\right)^l \zeta(l) + \frac{1}{\beta} \ln M\left(-\frac{\beta}{2}, 1; -\upsilon_k\right)\right\}. \tag{4.13}$$

Therefore,

$$\lim_{\beta\to 0} G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} e^{-\gamma/2} \lim_{\beta\to 0} \exp\left\{\frac{\ln M\left(-\frac{\beta}{2}, 1; -\upsilon_k\right)}{\beta}\right\} \tag{4.14}$$

$$= \frac{\sqrt{\upsilon_k}}{\gamma_k} e^{-\gamma/2} \exp\left\{\lim_{\beta\to 0} \frac{\frac{\partial}{\partial\beta} M\left(-\frac{\beta}{2}, 1; -\upsilon_k\right)}{M\left(-\frac{\beta}{2}, 1; -\upsilon_k\right)}\right\} \tag{4.15}$$

where L'Hopital's rule has been used. Using the power series definition of the confluent hypergeometric function in (3.45), it can be shown that

$$\lim_{\beta\to 0} M\left(-\frac{\beta}{2}, 1; -\upsilon_k\right) = 1 \tag{4.16}$$

and also (see [28]) that

$$\lim_{\beta\to 0} \frac{\partial}{\partial\beta} M\left(-\frac{\beta}{2}, 1; -\upsilon_k\right) = -\frac{1}{2}\sum_{r=1}^{\infty} \frac{(-\upsilon_k)^r}{r!}\frac{1}{r}. \tag{4.17}$$

Substituting (4.16) and (4.17) in (4.15), we obtain:

$$\lim_{\beta\to 0} G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \exp\left\{-\frac{\gamma}{2} - \frac{1}{2}\sum_{r=1}^{\infty} \frac{(-\upsilon_k)^r}{r!}\frac{1}{r}\right\}. \tag{4.18}$$

Using (8.214.1) from [83] which states that

$$-\int_a^{\infty} \frac{e^{-t}}{t} dt = \gamma + \ln(a) + \sum_{r=1}^{\infty} \frac{(-a)^r}{r!}\frac{1}{r}, \quad a > 0 \tag{4.19}$$

(4.18) becomes:

$$\lim_{\beta \to 0} G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \exp\left\{ \frac{1}{2} \left(\ln(\upsilon_k) + \int_{\upsilon_k}^{\infty} \frac{e^{-t}}{t} dt \right) \right\} \tag{4.20}$$

$$= \frac{\upsilon_k}{\gamma_k} \exp\left\{ \frac{1}{2} \int_{\upsilon_k}^{\infty} \frac{e^{-t}}{t} dt \right\} \tag{4.21}$$

which is the LSA gain as given by (3.40).

4.4 Concluding remarks

In this chapter, we extended the analysis of the β-SA estimator, originally presented in [29] for the case $\beta > 0$, to negative values of β, i.e. $\beta > -2$. We showed that negative values of β had a normalization effect on the original β-SA cost function. Moreover, decreasing β below 0 was found to produce an increase in the noise reduction and speech distortion, therefore enabling an extension of the trade-off between speech distortion and noise reduction. Finally, we proved mathematically that the case $\beta \to 0$ indeed corresponds to the LSA estimator.

It will be seen in Section 8.3 that the β-SA estimator with $\beta = -1$ slightly outperforms the well known MMSE STSA and LSA estimators in terms of the Perceptual Evaluation of Speech Quality (PESQ) and that the corresponding overall informal Mean Opinion Score (MOS) is found to be better than both MMSE STSA and LSA for white noise.

It is to be noted that there might be alternate closed-form solutions for $\beta \leq -2$. However, based on the experimental results in Section 8.3, it would seem that those values would not be interesting for speech enhancement. In fact, the speech distortion increases as the value of β decreases and values of $-2 < \beta < -1.5$ already produce much speech

distortions.

The work in Section 4.2 was presented in [99], while the proof in Section 4.3 appeared as a part of [100].

Chapter 5

Weighted β-SA estimator with auditory-based parameter values

As described in Chapter 3, the WE estimator [6] incorporates a weighting factor while the β-SA estimator [29] incorporates a power law. The parameters accounting for these effects can be given perceptual interpretations that were not considered in [6, 29]. In this chapter, we first derive and analyze a new family of Bayesian STSA estimators that combines the power law of the β-SA estimator and the weighting factor of the WE estimator, which we call the weighted β-SA family of estimators (Wβ-SA). We then present an original frequency dependent selection of the corresponding parameters based on perceptual considerations.

In Section 5.1, we briefly review the assumptions and cost functions of the β-SA and WE estimators and provide some motivation for the work performed in this chapter. In Section 5.2 we derive and analyze the proposed Wβ-SA family of estimators while in Section 5.3, the choice of perceptually significant parameter values is discussed.

5.1 Problem formulation and motivation

In this chapter we use the additive noise model in the STFT domain described in Section 3.1.2. The noisy speech is given by $Y_k = X_k + W_k$ where X_k and W_k denote the STFT coefficients of the clean speech and noise respectively. As in Section 3.4, we wish to find the estimator $\hat{X}_k^o = \hat{\mathcal{X}}_k^o e^{j\measuredangle Y_k}$ where $\hat{\mathcal{X}}_k^o$ is the estimator of the STSA coefficient of X_k and $\measuredangle Y_k$ is the phase of Y_k. In the Bayesian formalism $\hat{\mathcal{X}}_k^o$ is obtained as the minimum solution to $E\{C(\mathcal{X}_k, \hat{\mathcal{X}}_k)\}$.

A possible avenue for choosing an appropriate cost function $C(\mathcal{X}_k, \hat{\mathcal{X}}_k)$, and one that we will explore in this chapter, is to consider the human hearing mechanism. As discussed in Subsection 3.4.1, the MMSE STSA estimator is obtained when the chosen cost function is the squared error between the estimated and actual clean speech STSA [4]. Based on the assumption that the human auditory system performs a compression of the speech signal amplitude [85], the MMSE of the logarithm of the STSA (MMSE log-STSA or LSA) is proposed in [28]. Moreover, in [101, 102], masking thresholds are introduced in the Bayesian estimator's cost function to make it more perceptually significant while in [6], several perceptually relevant distortion metrics are considered as cost functions.

One of the estimators which was found to yield the best result in [6] is based on a perceptually-weighted error criterion used in speech coding. In this approach, the error spectrum is weighted by a filter which is the inverse of the original speech spectrum. To this end a generalization to the MMSE STSA cost function is proposed [6], i.e.:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \mathcal{X}_k^p \left(\mathcal{X}_k - \hat{\mathcal{X}}_k\right)^2 \tag{5.1}$$

which is the WE cost function defined in (3.47) and repeated here for convenience. In (5.1),

the parameter p controls the trade-off between speech distortion and noise reduction; the solution for the corresponding WE estimator is valid for $p > -2$ [6].

The β-SA estimator is another generalization of the MMSE STSA. Its cost function, repeated below for convenience,

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = (\mathcal{X}_k^\beta - \hat{\mathcal{X}}_k^\beta)^2 \tag{5.2}$$

and associated gain were studied in detail in Chapter 4. Based on this analysis, in particular, the range of the parameter β is taken as $\beta > -2$. While it was not interpreted as such in [29], the exponent β can be seen as performing a nonlinear compression on the STSA. It is known that the human ear performs a dynamic range compression [54] and, in fact, power laws have been used in auditory models to account for that compression [103].

To take advantage of the perceptual or auditory interpretation that can be given to both the parameters p and β, we propose here a new family of Bayesian STSA estimators that combines the weighting factor of the WE estimator and the power law of the β-SA estimator which we call the Weighted β-SA family of estimators (Wβ-SA). Moreover, we propose appropriate frequency dependent values for the parameters entering in the Wβ-SA cost function, i.e. β and α (the latter is related to the WE estimator parameter p). They are based on characteristics of the human auditory system among which are the compressive nonlinearities of the cochlea, the perceived loudness and the ears masking properties. The experimental evaluation of the new estimator will be addressed in Chapter 8.

5.2 The Wβ-SA family of estimator

In this section, we derive the Wβ-SA family of estimator. To do so, we seek to combine the β-SA and WE cost functions (respectively (5.2) and (5.1)) into a single cost function.

5.2.1 Derivation of the Wβ-SA estimator

The proposed cost function combining the desirable features of the cost functions in β-SA and WE is:

$$C(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \left(\frac{\mathcal{X}_k^\beta - \hat{\mathcal{X}}_k^\beta}{\mathcal{X}_k^\alpha} \right)^2 \tag{5.3}$$

where we used $\alpha = -p/2$ for convenience and β and α are real parameters whose ranges are discussed below.

By using (5.3) in (3.26), repeated here for convenience:

$$\hat{\mathcal{X}}_k^o = \arg\min_{\hat{\mathcal{X}}_k} \int_0^\infty C(\mathcal{X}_k, \hat{\mathcal{X}}_k) f_{\mathcal{X}_k|Y_k}(\mathcal{X}_k|Y_k) d\mathcal{X}_k \tag{5.4}$$

we obtain:

$$\hat{\mathcal{X}}_k^o = \left(\frac{E\{\mathcal{X}_k^{\beta-2\alpha}|Y_k\}}{E\{\mathcal{X}_k^{-2\alpha}|Y_k\}} \right)^{\frac{1}{\beta}}. \tag{5.5}$$

Using the Gaussian statistical model introduced in Section 3.2 where the clean speech and noise STFT coefficients are i.i.d. complex circular Gaussian random variables with zero mean, we know from [28] and Appendix A in [6] that:

$$E\{\mathcal{X}_k^m|Y_k\} = \frac{\upsilon_k^{m/2}}{\gamma_k^m} \Gamma\left(\frac{m}{2}+1\right) M\left(-\frac{m}{2}, 1; -\upsilon_k\right) |Y_k|^m \tag{5.6}$$

where $m > -2$. As in previous chapters, $\Gamma(a)$ and $M(a, b; z)$ are the gamma and confluent hypergeometric function, respectively defined in (3.44) and (3.45), and γ_k and υ_k are defined in (3.36). Using (5.6) in (5.5) with the appropriate values of the parameter m (i.e. $m = \beta - 2\alpha$ for the numerator and $m = -2\alpha$ for the denominator), we can show that:

$$\hat{\mathcal{X}}_k^o = G_k |Y_k|$$

where

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k}\left(\frac{\Gamma\left(\frac{\beta}{2}-\alpha+1\right)M\left(\alpha-\frac{\beta}{2},1;-\upsilon_k\right)}{\Gamma\left(-\alpha+1\right)M\left(\alpha,1;-\upsilon_k\right)}\right)^{1/\beta} \tag{5.7}$$

for $\beta > 2(\alpha-1)$, $\alpha < 1$. We will denote this new family of estimators as the Weighted β-SA (Wβ-SA). To ensure that this estimator corresponds in fact to a minimum of $E\{C(\mathcal{X}_k, \hat{\mathcal{X}}_k)\}$ we verified that the second derivative of the integral in (5.4) at $\hat{\mathcal{X}}_k^o$ is indeed positive, which is the case.

5.2.2 Analysis of the Wβ-SA estimator

Gain versus instantaneous SNR

The Wβ-SA estimator gain, G_k, in (5.7) depends on the parameters of the cost function (i.e. β and α) as well as on the *a posteriori* SNR, γ_k, and the *a priori* SNR, ξ_k, since υ_k is a function of γ_k and ξ_k (see (3.36)). Fig. 5.1 and Fig. 5.2 present gain curves as a function of the instantaneous SNR, $\gamma_k - 1$, for several β and α values and for $\xi_k = 0$ dB and $\xi_k = 10$ dB respectively.

As can be observed, the estimator's gain decreases when α increases and increases when β increases. The gain of the Wβ-SA estimator thus behaves similarly as that of the β-SA and WE estimators. Since the proposed estimator generalizes both the β-SA and WE estimators, the gains of the later can, in fact, be obtained by setting $\alpha = 0$ for β-SA and $\beta = 1$, $\alpha = -p/2$ for WE.

High instantaneous SNR gain

It was shown in [6] that the WE estimator tends to a Wiener estimator as the instantaneous SNR, $\gamma_k - 1$, tends to infinity. In fact, the more general Wβ-SA estimator also tends to a Wiener filter in that case.

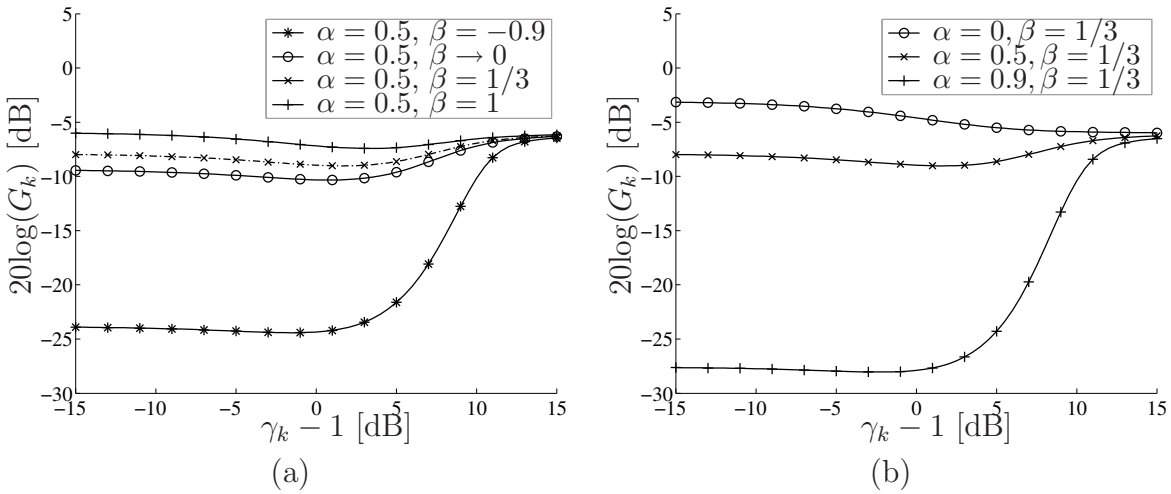

Fig. 5.1 Wβ-SA estimator gain ($20\log(G_k)$) versus instantaneous SNR, $\gamma_k - 1$, for $\xi_k = 0$ dB and (a) $\alpha = 0.5$ and $\beta \in \{-1, \to 0, 1/3, 1\}$ (b) $\alpha \in \{0, 0.5, 0.9\}$ and $\beta = 1/3$.

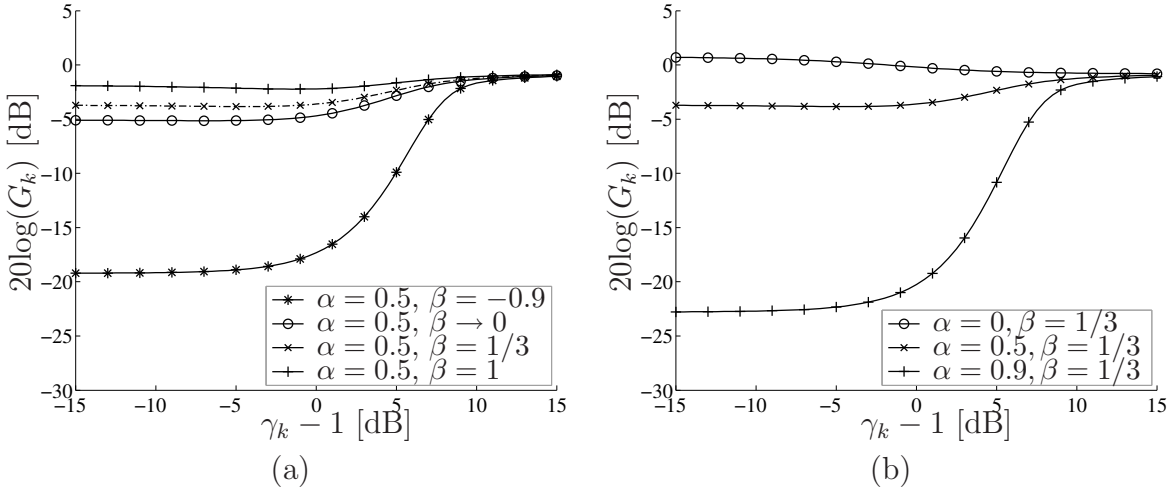

Fig. 5.2 Wβ-SA estimator gain ($20\log(G_k)$) versus instantaneous SNR, $\gamma_k - 1$, for $\xi_k = 10$ dB and (a) $\alpha = 0.5$ and $\beta \in \{-1, \to 0, 1/3, 1\}$ (b) $\alpha \in \{0, 0.5, 0.9\}$ and $\beta = 1/3$.

As $\gamma_k \to \infty$, we have from (3.36) that $\upsilon_k \to \infty$ and from (13.1.5) in [104] that:

$$\lim_{\upsilon_k \to \infty} M(-\frac{m}{2}, 1; -\upsilon_k) = \frac{\upsilon_k^{m/2}}{\Gamma(\frac{m}{2} + 1)}. \tag{5.8}$$

In the limit $\gamma_k \to \infty$, using (5.8) in (5.7) with the appropriate values of the parameter m, we thus have:

$$\lim_{\gamma_k \to \infty} G_k = \frac{\xi_k}{1+\xi_k} \tag{5.9}$$

which is a Wiener filter gain. Interestingly, since a Wiener filter results from the MMSE estimator of the STFT components as noted in Section 3.3, the use of the Wβ-SA cost function, as that of the WE and β-SA, is therefore equivalent to $C(X_k, \hat{X}_k) = (X_k - \hat{X}_k)^2$ for a high instantaneous SNR.

5.3 Choosing the β and α values based on auditory considerations

The parameter values of speech enhancement algorithms have traditionally been chosen based on frame SNR such that a higher gain is obtained for higher SNR and vice versa [1,29]. This had the effect of removing less noise at higher SNR to prevent speech distortion and more noise at low SNR.

Rather than focusing on the frame's SNR, we choose to consider the human auditory system to motivate the selection of appropriate values for β and α; in which case β and α will be fixed for all frames. In the first part of this section, we will present two different choices for β according to, firstly, the perceived loudness of sound and, secondly, the compressive nonlinearities of the cochlea. In the second part of this section, we will choose values of α considering the masking properties of the human auditory system. These different values will be compared through experimental results in Section 8.4 to assess their relevance in speech enhancement.

5.3.1 Choosing appropriate β values

A) Loudness considerations

In the LSA estimator, the logarithm of the spectral amplitude was considered. This was based on the fact that the MMSE of the logarithm of the spectral amplitude was thought to be more perceptually relevant than the spectral amplitude itself. By its very definition [54], loudness is more perceptually relevant than the sound's intensity. Therefore, a cost function which would consider the difference in terms of the perceived loudness would seem preferable to cost functions which consider the difference in terms of the sound intensity. Power laws have been used in the past to model the nonlinear relation between the intensity of sound and its perceived loudness [105, 106]. An exponent of 1/3 (i.e. cubic root) has been used in [106] to approximate the nonlinear transformation between intensity and perceived loudness. An appropriate value for the exponent β in the Wβ-SA cost function (5.3) would therefore be $\beta = 1/3$. This value will be further assessed experimentally in Section 8.4.

B) Compressive nonlinearities

An important factor that plays a role in the perception of loudness is the dynamic range compression performed by the ear [54]. This dynamic range compression is thought to be due to many factors among which are the cochlea's compressive nonlinearities. As mentioned in Subsection 2.2.2, compression rates of approximately 0.2 dB/dB were measured at the base of the mammalian cochlea (i.e. for high frequencies[1]) for intensities between 40 and 90 dB SPL [53][2]. These compression rates can be easily incorporated in the proposed Bayesian cost function (5.3). In fact, β can be directly interpreted as a compression rate,

[1] The characteristic frequencies of the tones in the corresponding studies varied from 8 kHz to 33 kHz depending on the species [53].

[2] As a comparison, conversational speech is at 60 dB SPL.

in dB/dB, of the input spectral amplitudes and thus set to corresponding physiologically meaningful values. Therefore, instead of motivating the value of β strictly in terms of loudness perception, we can also look at the physiology of the cochlea, which can explain to some extent the loudness perception of the human auditory system, and propose other relevant values for this parameter.

The cochlea's compressive nonlinearities are well documented and accepted at high frequencies, however, there is no consensus on the degree of nonlinearity at lower frequencies, i.e. at the apex of the cochlea [53, 54]. There would seem to be less compression at lower frequencies than at higher frequencies. In fact, some research even fail to show any compression at low frequencies, i.e. they observe a compression rate of 1 dB/dB, or even show an expansion, i.e. a compression rate > 1 dB/dB [53]. Here, we will assume no compressive nonlinearity at the low frequency limit. Since the compression rates will be different at low and high frequencies, the values of β will therefore be frequency dependent, which will be denoted by adding the subscript k to β, i.e. β_k.

To propose adequate values for the β_k's, we need to define the cochlea's rate of compression for every frequency k. Since for low frequency we consider the absence of compressive nonlinearity, we will choose β_k at the low frequency limit as $\beta_{\text{low}} = 1$. As indicated previously, the compressive nonlinearity at high frequencies is thought to have a compression rate of approximately 0.2 dB/dB. For high frequencies, it therefore seems plausible to set the high frequency limit of the β_k value as $\beta_{\text{high}} = 0.2$.

Physiological experiments on the cochlear rate of compression at intermediate frequencies, i.e. between the apex and the base of the cochlea, are extremely scarce [54]. Therefore we propose to interpolate β_k for intermediate frequencies based on the following approach. We consider the fact that each frequency corresponds to a position on the basilar membrane following the so-called tonotopic mapping [53]. One such tonotopic mapping, proposed

in [107], is given by:

$$d_k = \frac{1}{\rho}\log_{10}\left(\frac{f_k}{A}+1\right), \quad k = 1\ldots\frac{N}{2}-1 \tag{5.10}$$

where d_k is the position on the basilar membrane in mm, $\rho = 0.06$ mm^{-1} is an empirical scaling constant that depends on the basilar length, and is thus specific to a species, and $A = 165.4$ Hz is a scaling parameter allowing for the frequency to be expressed in Hz [107]. Moreover, f_k, in (5.10), is the frequency in Hz corresponding to spectral component k, i.e. $f_k = kF_s/N$ where F_s is the sampling frequency set to 16 kHz in this study and N is the DFT size typically set to 512 here if no zero padding is used.

We will therefore consider the compression rate to vary linearly with respect to the position d_k on the basilar membrane, corresponding to frequency f_k as given by the tonotopic mapping. In fact, the compressive nonlinearity is thought to be caused by the active process of the outer hair cells and it is known that the hair cells follow a tonotopic organization where they are optimally sensitive to a particular frequency according to their position on the basilar membrane [51]. The complete set of β_k values are thus derived by linearly interpolating between β_{low} and β_{high} according to d_k:

$$\beta_k = d_k\frac{(\beta_{\text{high}}-\beta_{\text{low}})}{\frac{1}{\rho}\log_{10}\left(\frac{F_s}{2A}+1\right)}+\beta_{\text{low}}. \tag{5.11}$$

Fig. 5.3 represents the different values of β_k as a function of the frequency.

C) Discussion

In the first part of this section we proposed the use of an exponent value $\beta = 1/3$ as a simple model for approximating the nonlinear transformation between intensity and perceived

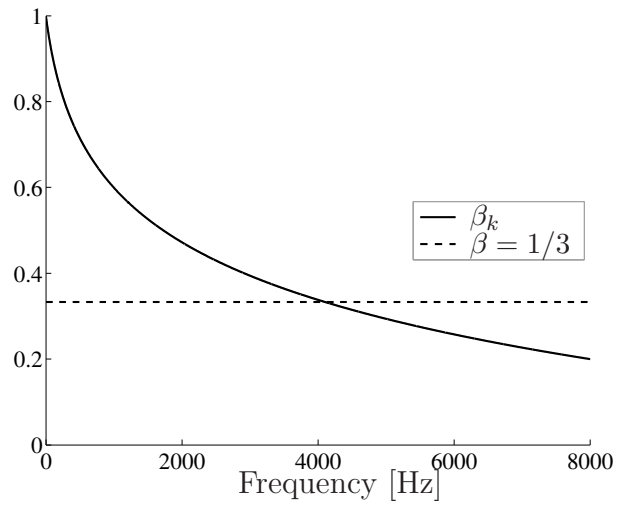

Fig. 5.3 Values of β_k and $\beta = 1/3$ versus frequency.

loudness. In the second part, we proposed frequency dependent β values based on the compressive nonlinearities of the cochlea. These are plotted in Fig. 5.3.

As can be seen from the gain curves in Fig. 5.1 and Fig. 5.2, the chosen values of β_k in Fig. 5.3, results in a decrease in the estimator's gain for high frequencies. This will therefore have for effect to limit the speech distortions at lower frequencies by keeping a higher gain and increasing the noise reduction at higher frequencies by decreasing the gain value.

It is interesting to note that more elaborate loudness models lead to a pattern of compression similar to the one described in the second part of this section. In fact, in the loudness model presented in [108], an exponent of 0.2 is used at high frequencies to perform compression while it is increased for lower frequencies.

5.3.2 Choosing appropriate α values

The WE estimator takes advantage of the masking properties of the human ear. In fact, one of the motivations for deriving the WE estimator was to favor a more accurate estimation

of smaller STSA since they are less likely to mask noise remaining in the clean speech estimate. This was done by choosing a fixed value of p that increased the weight of smaller STSA in the cost function (e.g. $p = -1$).

Since most of the speech energy is located at lower frequencies [109], higher frequencies should contain mainly small STSA. Therefore, it would be relevant to further increase the weights of the smaller STSA in the cost function for higher frequencies. This can be done by increasing α for higher frequencies, or equivalently decreasing p since $\alpha = -p/2$. We therefore propose, instead of using a fixed value of α as in [6], to modify α as a function of frequency, which we will denote by α_k, increasing its value for higher frequencies.

To do so, we need to choose appropriately the values of α_k for each frequency. In [6], the value of $p = -1$, corresponding here to $\alpha = 0.5$, has been suggested as a good compromise between the desired noise reduction performed by the estimator and the speech distortion introduced. This value can also be regarded as being a good compromise between increasing the weight of smaller STSA while keeping an appropriate estimation error for larger STSA. Since the main part of the speech energy, which will contain most of the larger STSA, is approximately located below 2000 Hz [109] (which also includes most of the first two formants [50]), we will choose the value of $\alpha = 0.5$ up to 2000 Hz. For higher frequencies, we want to further increase the weights of smaller STSA. Since, on average, the total speech energy decreases as frequency increases, we therefore propose to linearly increase the value of α as a function of the frequency. The Wβ-SA estimator restricts α to $\alpha < 1$, based on experimentations, we choose $\alpha = 0.9$ as the high frequency limit. Choosing higher values (e.g. $\alpha = 0.99$) did not introduce significant noise reduction while unnecessarily distorting

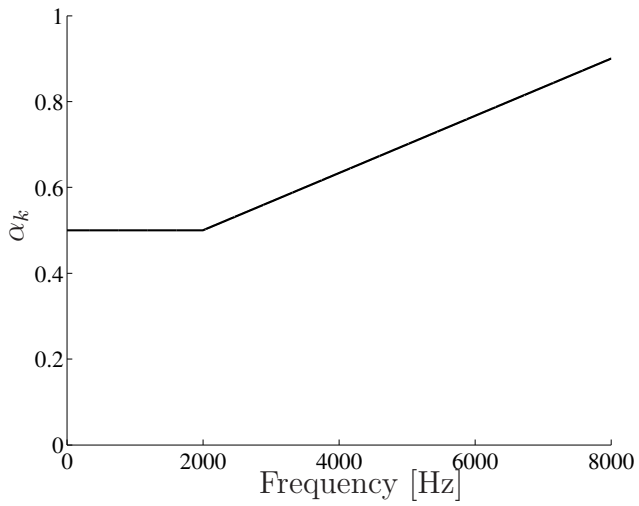

Fig. 5.4 Values of α_k versus frequency [Hz].

the speech. Therefore α_k will be given by:

$$\alpha_k = \begin{cases} \alpha_{\text{low}} & f_k \leq 2 \text{ kHz} \\ \frac{(f_k - 2000)(\alpha_{\text{high}} - \alpha_{\text{low}})}{F_s/2 - 2000} + \alpha_{\text{low}} & \text{else} \end{cases} \tag{5.12}$$

where $\alpha_{\text{low}} = 0.5$ and $\alpha_{\text{high}} = 0.9$. This relation is sketched in Fig. 5.4.

As can be seen again from the gain curves in Fig. 5.1 and Fig. 5.2, the chosen values of α_k in Fig. 5.4, results in a decrease in the estimator's gain for high frequencies, as does the β_k values.

5.4 Concluding remarks

In this chapter, we proposed a new family of Bayesian STSA estimators for speech enhancement, the Wβ-SA, where the cost function includes both a power law and a weighting factor. The corresponding estimator's gain parameters (i.e. β and α) were chosen according to characteristics of the human auditory system. It is found that choosing the parameters in

this way results in a decrease of the estimator gain at high frequencies. This frequency dependence of the gain improves the noise reduction while limiting the speech distortion.

Relevant experiments were conducted to evaluate the new Wβ-SA family of estimators with the proposed choices for β and α; the results are presented in Section 8.4. In particular, it will be shown, using both objective and subjective performance measures, that the new estimators achieve better enhancement performance, especially at low SNR values, when compared to existing Bayesian STSA estimators such as the MMSE STSA, LSA and WE estimator.

The work in this Chapter appeared as a journal paper in [100] and was presented in part in [110] and [111].

Chapter 6

Analytical generalization of Bayesian STSA estimators

The different cost functions presented in the previous chapters all have a structure involving a weighted squared difference between a monotonic function of the estimated and actual clean speech STSA. In this chapter, we perform an analytical generalization of existing Bayesian STSA estimators and develop a general family of Bayesian STSA estimators. This will allow for a unification of several existing Bayesian STSA estimators and will also provide a better understanding of this class of estimators.

In Section 6.1, we briefly review the problem framework and expose the similarities between existing cost functions. In Section 6.2, we develop and present the Generalized Weighted family of STSA estimators (GWSA) and in Section 6.3, we investigate some of its features.

6.1 Similarities between Bayesian STSA estimators

In this section, we examine the different cost functions that have been proposed recently, in the context of Bayesian STSA estimation, in order to reveal their similar structure. First let us briefly review the problem formulation. We use again here the additive noise model in the STFT domain described in Section 3.1.2 where the noisy speech is given by $Y_k = X_k + W_k$; X_k and W_k denoting the STFT coefficients of the clean speech and noise respectively. We wish to find the estimator $\hat{X}_k^o = \hat{\mathcal{X}}_k^o e^{j\measuredangle Y_k}$ where $\measuredangle Y_k$ is the phase of Y_k and $\hat{\mathcal{X}}_k^o$ is the estimator of the STSA coefficient of X_k. The latter is obtained as the minimum solution to $E\{C(\mathcal{X}_k, \hat{\mathcal{X}}_k)\}$ where $C(\mathcal{X}_k, \hat{\mathcal{X}}_k)$ is a suitably chosen cost function.

The β-SA (see Subsection 3.4.3) and WE (see Subsection 3.4.4) estimators were proposed as generalizations of the MMSE STSA estimator. Their cost functions (respectively (3.41) and (3.47)) have the following forms[1]:

$$C_{\beta\text{-SA}}(\mathcal{X}_k, \hat{\mathcal{X}}_k) \triangleq (\mathcal{X}_k^\beta - \hat{\mathcal{X}}_k^\beta)^2 \tag{6.1}$$

$$C_{\text{WE}}(\mathcal{X}_k, \hat{\mathcal{X}}_k) \triangleq \mathcal{X}_k^p (\mathcal{X}_k - \hat{\mathcal{X}}_k)^2 \tag{6.2}$$

where $\beta > -2$ and $p > -2$ respectively. In Chapter 5, the WE and β-SA estimators were combined in the Wβ-SA estimator for which the cost function is given by (5.3), repeated here for convenience:

$$C_{W\beta\text{-SA}}(\mathcal{X}_k, \hat{\mathcal{X}}_k) \triangleq \left(\frac{\mathcal{X}_k^\beta - \hat{\mathcal{X}}_k^\beta}{\mathcal{X}_k^\alpha} \right)^2 . \tag{6.3}$$

where $\beta > 2(\alpha - 1)$, $\alpha < 1$.

A variant of the well-known Itakura-Saito distortion measure, the COSH measure, was

[1] In this chapter, for ease of reading, we identify each cost function by a subscript denoting the corresponding estimator.

proposed in [6] as a cost function for Bayesian STSA estimation (see Subsection 3.4.5). This cost function can be shown to have a similar structure as the cost functions enumerated above. In fact, we have from (3.50) that:

$$C_{\mathrm{COSH}}(\mathcal{X}_k, \hat{\mathcal{X}}_k) \triangleq \frac{1}{2}\left(\frac{\mathcal{X}_k}{\hat{\mathcal{X}}_k} + \frac{\hat{\mathcal{X}}_k}{\mathcal{X}_k}\right) - 1 = \frac{(\mathcal{X}_k - \hat{\mathcal{X}}_k)^2}{2\mathcal{X}_k\hat{\mathcal{X}}_k}. \tag{6.4}$$

Moreover, a generalization of the COSH cost function, the WCOSH, was also proposed in [6] and is given by (3.51) which is repeated here:

$$C_{\mathrm{WCOSH}}(\mathcal{X}_k, \hat{\mathcal{X}}_k) \triangleq \left(\frac{\mathcal{X}_k}{\hat{\mathcal{X}}_k} + \frac{\hat{\mathcal{X}}_k}{\mathcal{X}_k} - 1\right)\mathcal{X}_k^q. \tag{6.5}$$

where $q > -1$. The WCOSH cost function can also be expressed in a similar form as the previous cost functions. In fact, we can modify the WCOSH cost function in the following form:

$$C'_{\mathrm{WCOSH}}(\mathcal{X}_k, \hat{\mathcal{X}}_k) = C_{\mathrm{WCOSH}}(\mathcal{X}_k, \hat{\mathcal{X}}_k) - \mathcal{X}_k^q = \frac{(\mathcal{X}_k - \hat{\mathcal{X}}_k)^2}{\mathcal{X}_k^{1-q}\hat{\mathcal{X}}_k} \tag{6.6}$$

without any modification on the final estimator since the cost function will be minimized with respect to $\hat{\mathcal{X}}_k$ in (3.26) to obtain the Bayesian STSA estimator. In fact, the term $\mathcal{X}_k^q$ subtracted from $C_{\mathrm{WCOSH}}(\mathcal{X}_k, \hat{\mathcal{X}}_k)$ will contribute the constant term $E\{\mathcal{X}_k^q|Y_k\}$ to the Bayesian objective function in (3.26).

In all the cost functions presented above, a similar structure can be highlighted. It involves a squared difference between a monotonic power function of $\mathcal{X}_k$ and $\hat{\mathcal{X}}_k$, this difference being further weighted by a function of either $\mathcal{X}_k$ or $\hat{\mathcal{X}}_k$ or both.

6.2 GWSA family of estimators

In this section, we generalize the common structure of the cost functions highlighted above and derive the corresponding closed-form solution.

6.2.1 A generalized cost function

We propose the following cost function:

$$C_{\mathrm{GWSA}}(\mathcal{X}_k, \hat{\mathcal{X}}_k) = \left(\frac{\mathcal{X}_k^{\beta} - \hat{\mathcal{X}}_k^{\beta}}{\mathcal{X}_k^{\alpha} \hat{\mathcal{X}}_k^{\eta}} \right)^2 \tag{6.7}$$

where the squared difference $(\mathcal{X}_k^{\beta} - \hat{\mathcal{X}}_k^{\beta})^2$ is now normalized by both $\mathcal{X}_k^{-2\alpha}$ and $\hat{\mathcal{X}}_k^{-2\eta}$; β, α and η are all real parameters for which the domains will be specified in the next subsection. The Bayesian family of STSA estimators obtained by minimizing the cost function (6.7) will be called the Generalized Weighted family of STSA estimators (GWSA).

The MMSE STSA, LSA, COSH, WE, β-SA, Wβ-SA and WCOSH estimators will then be all particular cases of the GWSA family of estimators with parameter values α, β and η as given in the corresponding columns of Table 6.1. We note that, in contrast to the existing cost functions, the cost function in (6.7) features a new parameter, η, that acts only on the estimated clean speech STSA, $\hat{\mathcal{X}}_k$.

Table 6.1 GWSA parameter values (β, α and η) corresponding to several existing Bayesian STSA estimators.

Estimator	Reference	β	α	η
MMSE STSA	[4], Subsection 3.4.1	1	0	0
LSA	[28], Subsection 3.4.2	$\to 0$	0	0
COSH	[6], Subsection 3.4.5	1	0.5	0.5
β-SA	[29], Subsection 3.4.3	β	0	0
WE	[6], Subsection 3.4.4	1	$-p/2$	0
WCOSH	[6], Subsection 3.4.5	1	$(1-q)/2$	0.5
Wβ-SA	Section 5.2	β	α	0

6.2.2 Derivation of the GWSA family of estimators

The Bayesian estimator corresponding to the cost function in (6.7) is obtained by finding the $\hat{\mathcal{X}}_k$ that minimizes the expectation of that given cost function, i.e.

$$\hat{\mathcal{X}}_k^o = \arg\min_{\hat{\mathcal{X}}_k} E\{C_{\mathrm{GWSA}}(\mathcal{X}_k, \hat{\mathcal{X}}_k)\} \tag{6.8}$$

$$= \arg\min_{\hat{\mathcal{X}}_k} \int_0^\infty \left(\frac{\mathcal{X}_k^\beta - \hat{\mathcal{X}}_k^\beta}{\mathcal{X}_k^\alpha \hat{\mathcal{X}}_k^\eta} \right)^2 f_{\mathcal{X}_k|Y_k}(\mathcal{X}_k|Y_k) d\mathcal{X}_k \tag{6.9}$$

where (6.7) has been used. Evaluating the first derivative of the integral in (6.9) with respect to $\hat{\mathcal{X}}_k$ and setting the result equal to zero, we get:

$$(2\eta-\beta)\hat{\mathcal{X}}_k^{\beta-1-2\eta} E\{\mathcal{X}_k^{\beta-2\alpha}\big|Y_k\} + (\beta-\eta)\hat{\mathcal{X}}_k^{2\beta-1-2\eta} E\{\mathcal{X}_k^{-2\alpha}\big|Y_k\} - \eta\hat{\mathcal{X}}_k^{-1-2\eta} E\{\mathcal{X}_k^{2\beta-2\eta}\big|Y_k\} = 0. \tag{6.10}$$

We notice that (6.10) has a quadratic form in $\hat{\mathcal{X}}_k^{\beta}$:

$$\hat{\mathcal{X}}_k^{-1-2\eta}(a\hat{\mathcal{X}}_k^{2\beta} + b\hat{\mathcal{X}}_k^{\beta} + c) = 0 \tag{6.11}$$

where the constants

$$a = (\beta - \eta)E\{\mathcal{X}_k^{-2\alpha}\big|Y_k\} \tag{6.12}$$

$$b = (2\eta - \beta)E\{\mathcal{X}_k^{\beta-2\alpha}\big|Y_k\} \tag{6.13}$$

$$c = -\eta E\{\mathcal{X}_k^{2\beta-2\eta}\big|Y_k\}. \tag{6.14}$$

Eq. (6.11) has trivial solutions at $\hat{\mathcal{X}}_k^o = 0$ or $\hat{\mathcal{X}}_k^o \to \infty$, depending on the values of η and β, and two non-trivial solutions obtained as the roots of $a\hat{\mathcal{X}}_k^{2\beta} + b\hat{\mathcal{X}}_k^{\beta} + c$:

$$\hat{\mathcal{X}}_k^o = \left(\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}\right)^{1/\beta}. \tag{6.15}$$

We discard the trivial solutions, which are not interesting for the current problem, and consider the solutions in (6.15).

Using the Gaussian statistical model introduced in Section 3.2, where the clean speech and noise STFT coefficients are i.i.d. complex circular Gaussian random variables with zero mean, we know from [28] and Appendix A in [6] that:

$$E\big\{\mathcal{X}_k^m\big|Y_k\big\} = \frac{\upsilon_k^{m/2}}{\gamma_k^m}\Gamma\Big(\frac{m}{2}+1\Big)\,M\Big(-\frac{m}{2},1;-\upsilon_k\Big)|Y_k|^m \tag{6.16}$$

where $m > -2$. As in previous chapters, $\Gamma(a)$ and $M(a,b;z)$ are the gamma and confluent hypergeometric function, respectively defined in (3.44) and (3.45), and the SNR parameters γ_k and υ_k are defined in (3.36). Using (6.16) in (6.12) - (6.14) and solving for the two

non-trivial solutions in (6.11), we finally obtain the following expression for the Bayesian estimator corresponding to the proposed GWSA cost function (6.7):

$$\hat{\mathcal{X}}_k^o = G_k|Y_k| \tag{6.17}$$

$$G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k}\left(\frac{-b' \pm \sqrt{b'^2 - 4a'c'}}{2a'}\right)^{\frac{1}{\beta}} \tag{6.18}$$

where the parameters:

$$a' = (\beta - \eta)\Gamma(-\alpha + 1)M(\alpha, 1; -\upsilon_k) \tag{6.19}$$

$$b' = (2\eta - \beta)\Gamma\left(\frac{\beta}{2} - \alpha + 1\right)M\left(\alpha - \frac{\beta}{2}, 1; -\upsilon_k\right) \tag{6.20}$$

$$c' = -\eta\Gamma(\beta - \alpha + 1)M(\alpha - \beta, 1; -\upsilon_k). \tag{6.21}$$

From the restriction on m in (6.16), we have that $\alpha < 1$ and $\beta > 2(\alpha - 1)$, and since a' cannot be equal to 0 in (6.18), we also have that $\beta \neq \eta$. Moreover, we need $b'^2 - 4a'c' \geq 0$ to avoid complex gains. Similar restrictions may also apply to the term inside the parenthesis in (6.18)[2], depending on the value of β.

We evaluated the second derivative of the integral in (3.14) with respect to $\hat{\mathcal{X}}_k$ to verify that the solutions in (6.18) are minimums. The result showed that the positive sign solution of (6.18) is a minimum if $\beta > 0$ and the negative sign solution is a minimum if $\beta < 0$. The chosen value of β therefore determines which of the positive or negative sign solution of (6.18) is appropriate.

[2]In fact, e.g. for $\beta = 2$, we need to have $(-b' \pm \sqrt{b'^2 - 4a'c'})/2a' > 0$.

6.3 Study of the GWSA family of estimators

In this section, we perform an analysis of the GWSA gain (6.18). We first analyze the behavior of the gain when plotted versus the instantaneous SNR, $\gamma_k - 1$, and then derive its value for high instantaneous SNR.

6.3.1 Gain versus instantaneous SNR

The GWSA gain depends on the parameters of the cost function (i.e. β, α and η) as well as on the parameters common to the previous Bayesian STSA estimators, namely the *a posteriori* SNR γ_k and the *a priori* SNR ξ_k, previously defined in (3.36).

Fig. 6.1 presents gain curves as a function of the instantaneous SNR, $\gamma_k - 1$, for a fixed $\xi_k = 0$ dB while in Fig. 6.2, $\xi_k = 10$ dB. In Fig. 6.1 (a) and Fig. 6.2 (a) we set $\beta = 1$ and show the gain curves for several α and η values while in Fig. 6.1 (b) and Fig. 6.2 (b) we set $\alpha = 0$ and show the gain curves for several β and η values. The case where $\eta = 0$ corresponds to the Wβ-SA estimator for which similar gain curves can be found in Fig. 5.1 and Fig. 5.2.

As can be observed in Fig. 6.1 and Fig. 6.2, the gain decreases when α increases, increases when η increases and *generally* increases when β increases. In fact, for some parameter values, the gain rather decreases as β increases (e.g. $\alpha = 0$ and $\eta = 0.8$ in Fig. 6.1 (b) and Fig. 6.2 (b)). Contrary to the existing estimators discussed previously, which all lead to a similar set of gain curves, the GWSA family of estimators provides with more flexibility in terms of achievable gain curves through the parameter η. In fact, with carefully chosen parameters, a steeper transition from high to low instantaneous SNR (e.g. Fig. 6.1 (b) and Fig. 6.2 (b) with $\alpha = 0$, $\beta \to 0$, $\eta = 0.8$) or an increase in the gain between $\gamma_k - 1 = -5$ dB and 10 dB (e.g. Fig. 6.1 (b) and Fig. 6.2 (b) with $\alpha = 0$, $\beta = 0.79$,

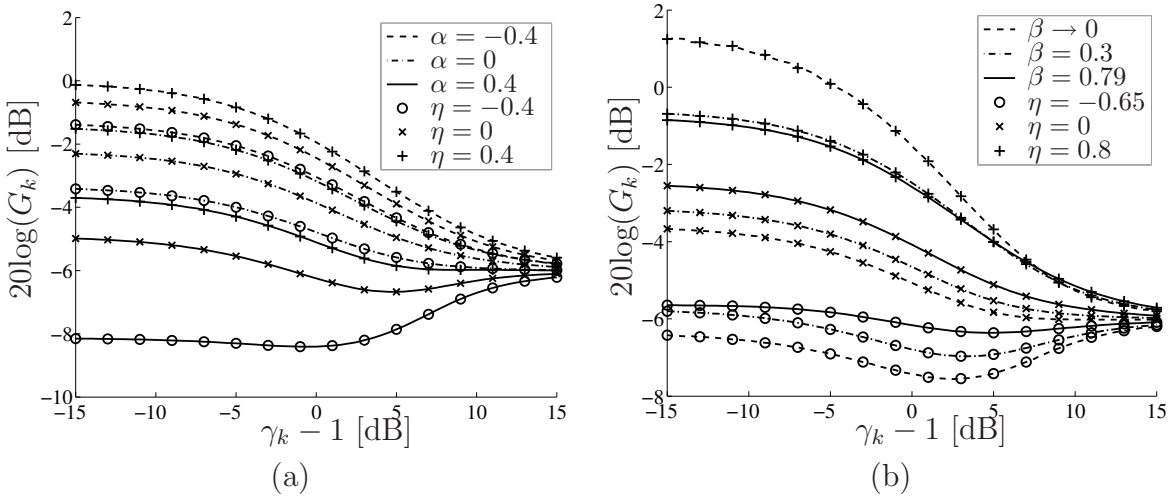

Fig. 6.1 GWSA estimator gain ($20\log(G_k)$) versus instantaneous SNR, $\gamma_k - 1$, with $\xi_k = 0$ dB for: (a) $\beta = 1$ and several α and η values; (b) $\alpha = 0$ and several β and η values.

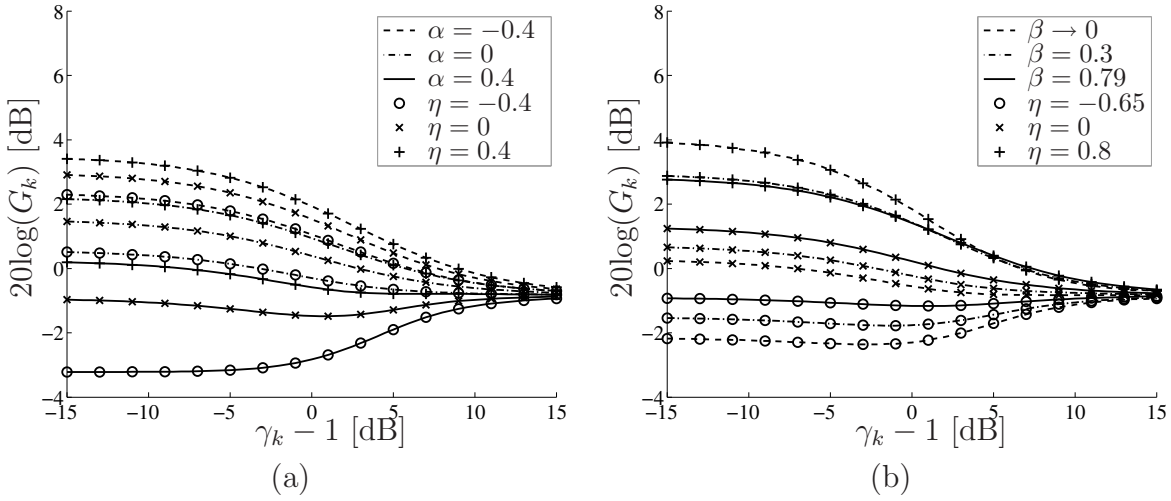

Fig. 6.2 GWSA estimator gain ($20\log(G_k)$) versus instantaneous SNR, $\gamma_k - 1$, with $\xi_k = 10$ dB for: (a) $\beta = 1$ and several α and η values; (b) $\alpha = 0$ and several β and η values .

$\eta = -0.65$) can be obtained. Care must be taken when choosing the parameters since an appropriate set of parameters for a given ξ_k may result in a complex gain for another value of ξ_k.

6.3.2 High instantaneous SNR gain

It was shown in Section 5.2 that the Wβ-SA estimator tends to the Wiener filter when the *a posteriori* SNR γ_k tends to infinity. In fact, all estimators belonging to the GWSA family converge to the Wiener filter in the limit $\gamma_k \to \infty$. Indeed, under this condition, we have from (3.36) that $\upsilon_k \to \infty$ and from from (13.1.5) in [104] that

$$\lim_{\upsilon_k \to \infty} M(-\frac{m}{2}, 1; -\upsilon_k) = \frac{\upsilon_k^{m/2}}{\Gamma(\frac{m}{2}+1)}. \tag{6.22}$$

In the limit $\gamma_k \to \infty$, using (6.22) in (6.19) - (6.21), (6.18) becomes:

$$\lim_{\gamma_k \to \infty} G_k = \frac{\sqrt{\upsilon_k}}{\gamma_k} \left[\upsilon_k^{\beta/2} \left(\frac{(\beta - 2\eta) \pm \sqrt{\beta^2}}{2\beta - 2\eta} \right) \right]^{\frac{1}{\beta}}. \tag{6.23}$$

Using (3.36), (6.23) for both the positive sign solution (with $\beta > 0$) and negative sign solution (with $\beta < 0$, i.e. $\beta = -|\beta|$) can easily be shown to simplify to:

$$\lim_{\gamma_k \to \infty} G_k = \frac{\xi_k}{1 + \xi_k} \tag{6.24}$$

which is the well-known Wiener filter gain.

All the members of the proposed GWSA family of Bayesian estimators, which include the β-SA, WE, WCOSH and Wβ-SA estimators, share the following common features:

- they span a wide range of gains at low instantaneous SNR depending on the different values of their respective parameters,
- they converge to the Wiener estimator for larger instantaneous SNR's.

While specific values of the η parameter could yield some advantage in terms of speech

enhancement in some environments, the main contribution of this chapter is more in the unification of the different Bayesian STSA estimators and the theoretical analysis that can be derived from it.

6.4 Concluding remarks

In this chapter, we first noted that several existing Bayesian STSA cost functions for speech enhancement are similarly structured. We therefore proposed an analytical generalization of the corresponding estimators as the GWSA family of estimators. The latter incorporates the parameters present in other existing estimators (e.g. α and β) but also features a new parameter: η. These parameters control the shape of the estimator's gain curve as a function of the instantaneous SNR. In contrast to the other parameters, η acts only on the estimated clean speech STSA. It is found that, for appropriate parameter values, η yields an added flexibility in terms of achievable gain curves when compared to existing Bayesian STSA estimators. Finally, we also showed that all the estimators belonging to the new estimator family tend to a Wiener filter for high instantaneous SNR. This work thus allowed a unification of several existing Bayesian STSA estimators and, moreover, also provided a better understanding of this general class of estimators. The work in this chapter appeared in [112].

Chapter 7

Multi-dimensional Bayesian STSA estimators allowing correlated frequency components

In the traditional Bayesian STSA estimation approach used in the previous chapters, the spectral components are assumed uncorrelated. However, this assumption is inexact since some correlation is present in practice. In this chapter, we investigate a multi-dimensional Bayesian STSA estimator that assumes correlated frequency components in digitized speech. Since the closed-form solution of this optimum estimator is not readily available, we alternatively derive closed-form expressions for an upper and a lower bound on the desired estimator. Using these bounds, we propose a new family of speech enhancement estimators.

In Section 7.1, we provide some motivation for the estimators developed in this chapter. In Section 7.2 we elaborate on the correlation that exists between the frequency components of digitized speech. Section 7.3 presents the desired multi-dimensional STSA estimator that

allows for correlated frequency components and develops the above mentioned lower and upper bounds as well as the proposed family of estimators. Section 7.4 studies the proximity between the upper and lower bounds and addresses the estimation of the associated correlation matrices.

7.1 Motivation

In the previous chapters, we considered an additive noise model

$$Y_k = X_k + W_k. \tag{7.1}$$

for a particular frame where Y_k, X_k and W_k denoted the STFT of the noisy speech, clean speech and noise respectively. We also defined $X_k = \mathcal{X}_k e^{j\theta_k}$ where $\mathcal{X}_k > 0$ is the STSA and $\theta_k \in [-\pi, \pi)$ is the associated phase. The goal in that traditional approach used in the previous chapters is then to obtain the estimator of the STSA, $\hat{\mathcal{X}}_k$, as a function of the noisy observations Y_k, which minimizes the expectation of a given cost function $C(\mathcal{X}_k, \hat{\mathcal{X}}_k)$:

$$\hat{\mathcal{X}}_k^o = \underset{\hat{\mathcal{X}}_k}{\operatorname{argmin}}\, E\{C(\mathcal{X}_k, \hat{\mathcal{X}}_k)\}. \tag{7.2}$$

This estimator is then combined with the phase of the noisy speech, $\measuredangle Y_k$, to yield the estimator of the complex STFT $\hat{X}_k = \hat{\mathcal{X}}_k e^{j\measuredangle Y_k}$.

The estimators of the STSA obtained through (7.2) have shown some advantages over estimators of the STFT such as the well known Wiener filter (3.22) [37]. In fact, one desirable feature of Bayesian STSA estimators, when combined with the *decision-directed* approach for the estimation of ξ_k, is to produce a residual background noise that is whiter than the residual musical noise produced by the traditional Wiener estimator [88].

In traditional Bayesian STSA estimation approaches, it is always assumed that the different spectral components of the clean speech STFT are uncorrelated so that the different frequency components of the noisy speech can be processed independently as in (7.2). This assumption is however inexact as there are some sources of correlation between the spectral components [113]. Firstly, the fundamental frequency of voiced speech has harmonics that are inherently correlated. Secondly, the finite temporal extension of the analysis window used in short-time processing introduces some correlation between adjacent frequencies.

A multi-dimensional MMSE estimator of the complex STFT coefficients that assumes correlated frequency components has recently been studied in [113]. This work focuses on obtaining an accurate estimation of the clean speech correlation matrix which is required in the solution of the underlying MMSE estimation problem. The resulting estimator is shown to be advantageous over several existing estimators, including a Wiener filter assuming uncorrelated frequency components.

On the one hand, Bayesian estimators of the STSA have been found to yield some advantages over Bayesian estimators of the complex STFT components. On the other hand, STFT estimators considering correlated frequency components yield better results than estimators not considering such correlation. Therefore, it appears that the consideration of correlated frequency components in Bayesian STSA estimation might lead to even superior performance. However, this avenue has apparently not been considered in the recent speech and audio literature.

In this chapter, we first investigate a multi-dimensional Bayesian STSA estimator that considers the spectral components of digitized speech to be correlated. Since a closed-form solution for such an estimator is not readily available, we alternatively develop closed-form expressions for a lower and an upper bound on the desired estimator. Based on those bounds, we propose a family of speech enhancement estimators being characterized by a

scalar parameter $0 \leq \gamma \leq 1$, with $\gamma = 0$ corresponding to the lower bound and $\gamma = 1$ to the upper bound. Knowledge of the clean speech and noise correlation matrices is needed to implement the proposed estimators. Since digitized speech has some correlation that is present only in voiced parts, we also modify the clean speech correlation matrix to give it a full structure in voiced sections and a diagonal structure in unvoiced sections.

The following notation is used in this chapter: for any vector $\mathbf{A} = [a_k] \in \mathbb{R}^{N\times 1}$ and any positive real b, we define $\mathbf{A}^{[b]} = [a_k^b]$; for any vector $\mathbf{A} \in \mathbb{C}^{N\times 1}$, we define $|\mathbf{A}| = [|a_k|]$; for any matrix $\mathbf{A} \in \mathbb{C}^{N\times N}$ we define $\text{diag}\{\mathbf{A}\}$ as the column vector containing the diagonal elements of matrix $\mathbf{A}$; $\mathbf{I}_N$ is the $N \times N$ identity matrix.

7.2 Correlation between the frequency components

In practice, and in contrast with the traditional assumptions used in the development of the estimators presented in the previous chapters, there is correlation between the different short-time spectral components of a speech signal. This correlation is due to different factors including:

- *Use of window in frame-based processing*: Indeed, the use of a finite analysis window function $h_a[n]$ in the computation of the STFT in (3.3) introduces some correlation between adjacent frequency components. This is due to the spectral smearing phenomenon which is a known effect of the windowing process [49].

- *Harmonic structure of voiced speech*: Voiced speech is characterized by the vibration of the vocal cords at a fundamental frequency F0 and has several harmonics at multiples of F0 [50] (see Section 2.1 for a brief review of the human speech production system). The frequencies corresponding to these different harmonics will therefore be inherently correlated.

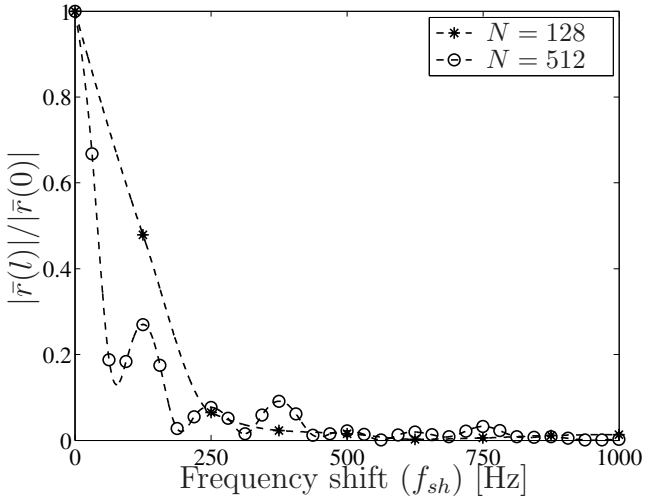

Fig. 7.1 Mean sample autocorrelation function $|\bar{r}(l)|/|\bar{r}(0)|$ versus the physical frequency shift $f_{sh} = lF_s/N$ in Hz for the vowel part of the male spoken word "hood".

To illustrate some of the correlation that exists between the spectral components in a speech signal, we conducted some experiments using several utterances at a sampling frequency of $F_s = 16000$ Hz, using two different frame lengths ($N = 128$, $N = 512$) and a 75% overlap between frames. For each frame i, the following sample correlation function was computed:

$$r_i(l) = \frac{1}{N-l} \sum_{k=0}^{N-1-l} (X_{k,i} - \bar{X}_i)(X_{k+l,i} - \bar{X}_i)^* \tag{7.3}$$

where $l \in \{0, 1, \ldots, N-1\}$ and $\bar{X}_i = \frac{1}{N} \sum_{k=0}^{N-1} X_{k,i}$ is the sample mean. The mean sample correlation

$$\bar{r}(l) = \frac{1}{N_f} \sum_{i=0}^{N_f-1} r_i(l) \tag{7.4}$$

where N_f is the total number of frames in the studied utterance was further evaluated.

In Fig. 7.1, we show $|\bar{r}(l)|/|\bar{r}(0)|$ versus the physical frequency shift $f_{sh} = lF_s/N$ in Hz for the vowel part of the male spoken word "hood" obtained from [114]. We observe that correlation is higher for lower values of l, i.e. between nearby frequencies. In the case

$N = 512$ (for which the resolution in frequency is finer) we can also observe correlations at multiples of ~ 125 Hz, which corresponds to the fundamental frequency of the speaker in this experiment. We also noticed, through a different set of experiments not shown here, that evaluating correlations for the STSA instead of the STFT yielded an even higher degree of correlation.

Based on these observations, we see that the different spectral components are indeed correlated. The estimators following the traditional uncorrelated approach are thus sub-optimal.

7.3 Family of multi-dimensional STSA estimators allowing correlated frequency components

In this section, we proceed to obtain a multi-dimensional clean speech STSA estimator that assumes correlated frequency components. Defining $\mathbf{Y} = [Y_0\ Y_1\ \cdots\ Y_{N-1}]^T$, it follows from (7.1) that:

$$\mathbf{Y} = \mathbf{X} + \mathbf{W}, \tag{7.5}$$

where $\mathbf{X} = [X_0\ X_1\ \cdots\ X_{N-1}]^T$ and $\mathbf{W} = [W_0\ W_1\ \cdots\ W_{N-1}]^T$ are respectively the clean speech vector and the noise vector of the corresponding STFT coefficients. We also define the STSA vector $\boldsymbol{\mathcal{X}} = [\mathcal{X}_0\ \mathcal{X}_1\ \cdots\ \mathcal{X}_{N-1}]^T$ and the phase vector $\boldsymbol{\theta} = [\theta_0\ \theta_1\ \cdots\ \theta_{N-1}]^T$. We assume that $\mathbf{X}$ and $\mathbf{W}$ are independent, zero-mean and circular Gaussians with probability density functions:

$$f_{\mathbf{X}}(\mathbf{X}) = \frac{1}{\pi^N \det(\mathbf{R}_{\mathbf{X}})} e^{-\mathbf{X}^H \mathbf{R}_{\mathbf{X}}^{-1} \mathbf{X}}, \tag{7.6}$$

$$f_{\mathbf{W}}(\mathbf{W}) = \frac{1}{\pi^N \det(\mathbf{R}_{\mathbf{W}})} e^{-\mathbf{W}^H \mathbf{R}_{\mathbf{W}}^{-1} \mathbf{W}}. \tag{7.7}$$

In these expressions $\mathbf{R_X} = E\{\mathbf{X}\mathbf{X}^H\}$ and $\mathbf{R_W} = E\{\mathbf{W}\mathbf{W}^H\}$ are the correlation matrices of the clean speech and of the noise respectively, superscript H indicates the conjugate transpose and $\mathbf{R_W} > 0$ (positive definite) is assumed. In the previous chapters, it is assumed that $\mathbf{R_X}$ and $\mathbf{R_W}$ are diagonal matrices, i.e. the spectral components are uncorrelated. In this chapter, we do not enforce such diagonality constraint. Our model therefore considers possible frequency correlations in the clean speech and noise.

We want to evaluate the MMSE estimator of $\boldsymbol{\mathcal{X}}$:

$$\hat{\boldsymbol{\mathcal{X}}}^o = \underset{\hat{\boldsymbol{\mathcal{X}}}}{\operatorname{argmin}}\, E\{\|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}\|^2\} \tag{7.8}$$

where the minimum is over all possible functions $\hat{\boldsymbol{\mathcal{X}}} \equiv \hat{\boldsymbol{\mathcal{X}}}(\mathbf{Y})$ of the observation vector $\mathbf{Y}$. We note that the cost function in (7.8), i.e. $C(\boldsymbol{\mathcal{X}}, \hat{\boldsymbol{\mathcal{X}}}) \triangleq \|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}\|^2$, considers all the STSA frequency components jointly. Using matrix calculus, we can show that (7.8) leads to:

$$\hat{\boldsymbol{\mathcal{X}}}^o = E\{\boldsymbol{\mathcal{X}}|\mathbf{Y}\} \tag{7.9}$$

i.e. the N-dimensional conditional expectation of $\boldsymbol{\mathcal{X}}$ given the complete vector of observations $\mathbf{Y}$. This estimator can then be combined with the phase of the noisy speech, for each frequency, to yield the estimator of $\mathbf{X}$:

$$\hat{\mathbf{X}}^o = [\hat{\mathcal{X}}_0^{o,s} e^{j\measuredangle Y_0},\ \cdots\ , \hat{\mathcal{X}}_{N-1}^{o,s} e^{j\measuredangle Y_{N-1}}]^T. \tag{7.10}$$

where superscript s is used to distinguish these STSA estimators from those obtained using (7.2).

Unfortunately and in contrast to the scalar case, a closed-form expression for (7.9) is not readily available. Since the $\hat{\mathcal{X}}_k^{o,s}$ are positive real quantities, we approach the problem

of finding a realizable approximation to (7.9) by first obtaining tractable upper and lower bounds, $\hat{\mathcal{X}}^o_{U,k}$ and $\hat{\mathcal{X}}^o_{L,k}$ respectively, such that $\hat{\mathcal{X}}^o_{L,k} < \hat{\mathcal{X}}^{o,s}_k < \hat{\mathcal{X}}^o_{U,k}$. Based on the obtained bounds, we will then propose a parameterized family of estimators.

7.3.1 Lower bound

Using the triangle inequality for integration [115], we can show that:

$$|E\{X_k|\mathbf{Y}\}| \leq E\{\mathcal{X}_k|\mathbf{Y}\}. \tag{7.11}$$

As a lower bound on the desired estimator (7.9), we therefore propose $\hat{\mathcal{X}}^o_{L,k} = |E\{X_k|\mathbf{Y}\}|$ or equivalently:

$$\hat{\boldsymbol{\mathcal{X}}}^o_L = |E\{\mathbf{X}|\mathbf{Y}\}| \tag{7.12}$$

where the $\hat{\mathcal{X}}^o_{L,k}$'s are the elements of the N-dimensional column vector $\hat{\boldsymbol{\mathcal{X}}}^o_L$. Under the Gaussian statistical model for the clean speech and noise presented previously, the term $E\{\mathbf{X}|\mathbf{Y}\}$ is the MMSE estimator of $\mathbf{X}$, which is known to be equal to [113]:

$$E\{\mathbf{X}|\mathbf{Y}\} = \hat{\mathbf{X}}_{\text{MMSE}} = \mathbf{G}_{\text{MMSE}}\mathbf{Y} \tag{7.13}$$

where the MMSE gain matrix $\mathbf{G}_{\text{MMSE}}$ is:

$$\mathbf{G}_{\text{MMSE}} \triangleq \mathbf{R}_{\mathbf{X}}(\mathbf{R}_{\mathbf{X}} + \mathbf{R}_{\mathbf{W}})^{-1}. \tag{7.14}$$

For future reference, it is also convenient to express $\mathbf{G}_{\text{MMSE}}$ in the following form, which can be obtained by the application of the matrix inversion lemma [116]:

$$\mathbf{G}_{\text{MMSE}} \triangleq (\mathbf{R_X}^{-1} + \mathbf{R_W}^{-1})^{-1}\mathbf{R_W}^{-1}. \tag{7.15}$$

A lower bound on the desired estimator is therefore:

$$\hat{\boldsymbol{\mathcal{X}}}_L^o = \left|\mathbf{G}_{\text{MMSE}}\mathbf{Y}\right|. \tag{7.16}$$

Note that in the special case of uncorrelated frequency components (i.e. the traditional framework), $\mathbf{R_X}$ and $\mathbf{R_W}$ in (7.14) are diagonal matrices. In that case, combining (7.16) with the phase of the noisy speech yields:

$$\hat{X}_k = \frac{\sigma_{X,k}^2}{\sigma_{X,k}^2 + \sigma_{W,k}^2} Y_k \tag{7.17}$$

where $\sigma_{X,k}^2 = [\mathbf{R_X}]_{kk} = E\{\mathcal{X}_k^2\}$ and $\sigma_{W,k}^2 = [\mathbf{R_W}]_{kk} = E\{|W_k|^2\}$. The processing of each frequency is therefore decoupled and the corresponding operation amounts to a standard Wiener filter as in (3.22).

7.3.2 Upper bound

Using Jensen's inequality [117], we have for a real convex function $\varphi(\cdot)$ that:

$$\varphi(E\{\mathcal{X}_k|\mathbf{Y}\}) \leq E\{\varphi(\mathcal{X}_k)|\mathbf{Y}\}. \tag{7.18}$$

If we set $\varphi(a) = a^2$, we obtain $E\{\mathcal{X}_k|\mathbf{Y}\}^2 \leq E\{\mathcal{X}_k^2|\mathbf{Y}\}$ and,

$$E\{\mathcal{X}_k|\mathbf{Y}\} \leq \sqrt{E\{\mathcal{X}_k^2|\mathbf{Y}\}} \tag{7.19}$$

which is also a special case of Lyapunov's inequality [38]. As an upper bound on the desired estimator (7.9), we therefore propose $\hat{\mathcal{X}}_{U,k}^o = \sqrt{E\{\mathcal{X}_k^2|\mathbf{Y}\}}$ or equivalently:

$$\hat{\boldsymbol{\mathcal{X}}}_U^o = E\{\boldsymbol{\mathcal{X}}^{[2]}|\mathbf{Y}\}^{[1/2]} \tag{7.20}$$

where the $\hat{\mathcal{X}}_{U,k}^o$'s are the elements of the N-dimensional column vector $\hat{\boldsymbol{\mathcal{X}}}_U^o$. We next derive a closed-form expression for $E\{\mathcal{X}_k^2|\mathbf{Y}\}$.

Using a Bayesian formalism we have:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \frac{\int\cdots\int |X_k|^2 f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{X}) f_{\mathbf{X}}(\mathbf{X}) d\mathbf{X}}{\int\cdots\int f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{X}) f_{\mathbf{X}}(\mathbf{X}) d\mathbf{X}}. \tag{7.21}$$

We observe from (7.5) that:

$$f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{X}) = f_{\mathbf{W}}(\mathbf{Y} - \mathbf{X}). \tag{7.22}$$

Using (7.6), (7.7) and (7.22) in (7.21) we obtain:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \frac{\int\cdots\int |X_k|^2 e^{\{\mathbf{Y}^H \mathbf{R_W}^{-1}\mathbf{X} + \mathbf{X}^H \mathbf{R_W}^{-1}\mathbf{Y} - \mathbf{X}^H(\mathbf{R_W}^{-1} + \mathbf{R_X}^{-1})\mathbf{X}\}} d\mathbf{X}}{\int\cdots\int e^{\{\mathbf{Y}^H \mathbf{R_W}^{-1}\mathbf{X} + \mathbf{X}^H \mathbf{R_W}^{-1}\mathbf{Y} - \mathbf{X}^H(\mathbf{R_W}^{-1} + \mathbf{R_X}^{-1})\mathbf{X}\}} d\mathbf{X}} \tag{7.23}$$

To evaluate (7.23), we need to transform the multiple integrals into products of single integrals. To do so, we make use of the following eigenvalue decomposition:

$$\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H = \mathbf{R_W}^{-1} + \mathbf{R_X}^{-1} \tag{7.24}$$

where $\mathbf{U}$ is the unitary matrix of eigenvectors, i.e. $\mathbf{U}^H\mathbf{U} = \mathbf{I}_N$, and $\mathbf{\Lambda}$ is the diagonal matrix containing the corresponding eigenvalues. Furthermore, we perform the following change of variables: $\mathbf{V} = \mathbf{U}^H\mathbf{X}$. Since $\mathbf{U}$ is unitary, the associated Jacobian is equal to 1 and (7.23) thus becomes:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \frac{\int\cdots\int |\mathbf{U}_k\mathbf{V}|^2 e^{\{\tilde{\mathbf{Y}}^H\mathbf{V}+\mathbf{V}^H\tilde{\mathbf{Y}}-\mathbf{V}^H\mathbf{\Lambda}\mathbf{V}\}}d\mathbf{V}}{\int\cdots\int e^{\{\tilde{\mathbf{Y}}^H\mathbf{V}+\mathbf{V}^H\tilde{\mathbf{Y}}-\mathbf{V}^H\mathbf{\Lambda}\mathbf{V}\}}d\mathbf{V}} \tag{7.25}$$

where we define $\mathbf{U}_k$ as the k^{th} line of $\mathbf{U}$ and

$$\tilde{\mathbf{Y}} \triangleq \mathbf{U}^H{\mathbf{R}_\mathbf{W}}^{-1}\mathbf{Y}. \tag{7.26}$$

Since $\mathbf{U}_k\mathbf{V}$ is a scalar, we have:

$$\mathbf{U}_k\mathbf{V} = \sum_{r=0}^{N-1} U_{kr}V_r \tag{7.27}$$

where U_{kr} is the kr^{th} entry of matrix $\mathbf{U}$ and V_r is the r^{th} entry of vector $\mathbf{V}$. Using (7.27), we can now write (7.25) in a form comprising only scalars:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \frac{\sum_{r=0}^{N-1}\sum_{t=0}^{N-1} U_{kt}^*U_{kr}\int\cdots\int V_t^*V_r\prod_{m=0}^{N-1}[g(V_m)dV_m]}{\prod_{m=0}^{N-1}\int g(V_m)dV_m} \tag{7.28}$$

where we define the positive real scalar function $g(V_m) = e^{\tilde{Y}_m^*V_m+V_m^*\tilde{Y}_m-|V_m|^2\lambda_m}$ for compactness and λ_m is the mth diagonal element of matrix $\mathbf{\Lambda}$.

Using (6.631.1), (8.411.1) and (9.212.1) from [83], we can evaluate the integrals in (7.28)

and obtain (see Appendix A):

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \sum_{r=0}^{N-1}\sum_{t=0}^{N-1} U_{kt}^* U_{kr} \frac{\tilde{Y}_t^* \tilde{Y}_r}{\lambda_t \lambda_r} + \sum_{p=0}^{N-1} \frac{|U_{kp}|^2}{\lambda_p}. \tag{7.29}$$

This last equation can also be equivalently written as:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \mathbf{U}_k \mathbf{\Lambda}^{-1} \tilde{\mathbf{Y}} \tilde{\mathbf{Y}}^H \mathbf{\Lambda}^{-1} \mathbf{U}_k^H + \mathbf{U}_k \mathbf{\Lambda}^{-1} \mathbf{U}_k^H. \tag{7.30}$$

which, in turn, using the notation introduced previously, can be expressed in a more compact form as:

$$E\{\boldsymbol{\mathcal{X}}^{[2]}|\mathbf{Y}\} = \text{diag}\{\mathbf{U}\mathbf{\Lambda}^{-1}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^H\mathbf{\Lambda}^{-1}\mathbf{U}^H + \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^H\}. \tag{7.31}$$

Using (7.24) and (7.26) along with the fact that $\text{diag}\{\mathbf{A}\mathbf{A}^H\} = |\mathbf{A}|^{[2]}$ for any $\mathbf{A} \in \mathbb{C}^{N\times 1}$, we have:

$$E\{\boldsymbol{\mathcal{X}}^{[2]}|\mathbf{Y}\} = |(\mathbf{R_W}^{-1} + \mathbf{R_X}^{-1})^{-1}\mathbf{R_W}^{-1}\mathbf{Y}|^{[2]} + \text{diag}\{(\mathbf{R_W}^{-1} + \mathbf{R_X}^{-1})^{-1}\}. \tag{7.32}$$

In light of (7.15), we notice that the entries of the first term in (7.32) are equal to the squared magnitudes of the entries of $\hat{\mathbf{X}}_{\text{MMSE}}$ in (7.13) and that the second term is simply $\text{diag}\{\mathbf{G}_{\text{MMSE}}\mathbf{R_W}\}$. Finally, using (7.20), the desired upper bound is obtained as the following simple expression:

$$\hat{\boldsymbol{\mathcal{X}}}_U^o = (|\mathbf{G}_{\text{MMSE}}\mathbf{Y}|^{[2]} + \text{diag}\{\mathbf{G}_{\text{MMSE}}\mathbf{R_W}\})^{[1/2]}. \tag{7.33}$$

Since the upper bound includes the lower bound and an additional positive term, it will obviously be greater than the lower bound.

7.3.3 Proposed family of estimators

The true estimator $\hat{\mathcal{X}}_k^{o,s}$ is smaller than $\hat{\mathcal{X}}_{U,k}^o$ and greater than $\hat{\mathcal{X}}_{L,k}^o$. Based on the expressions of the derived bounds $\hat{\mathcal{X}}_{L,k}^o$ in (7.16) and $\hat{\mathcal{X}}_{U,k}^o$ in (7.33) we therefore propose the following family of estimators:

$$\hat{\boldsymbol{\mathcal{X}}}_\gamma^o = (|\mathbf{G}_{\text{MMSE}}\mathbf{Y}|^{[2]} + \gamma\,\text{diag}\{\mathbf{G}_{\text{MMSE}}\mathbf{R}_{\mathbf{W}}\})^{[1/2]} \tag{7.34}$$

where $0 \leq \gamma \leq 1$. We have that $\hat{\boldsymbol{\mathcal{X}}}_L^o \leq \hat{\boldsymbol{\mathcal{X}}}_\gamma^o \leq \hat{\boldsymbol{\mathcal{X}}}_U^o$ with the limit cases:

$$\hat{\boldsymbol{\mathcal{X}}}_\gamma^o = \begin{cases} \hat{\boldsymbol{\mathcal{X}}}_U^o & \text{if } \gamma = 1 \\ \hat{\boldsymbol{\mathcal{X}}}_L^o & \text{if } \gamma = 0. \end{cases} \tag{7.35}$$

As in (7.10), the spectral amplitude estimators $\hat{\boldsymbol{\mathcal{X}}}_L^o$, $\hat{\boldsymbol{\mathcal{X}}}_U^o$ and $\hat{\boldsymbol{\mathcal{X}}}_\gamma^o$ are then combined with the phase of the noisy speech to obtain the corresponding complex spectrum estimators $\hat{\mathbf{X}}_L^o$, $\hat{\mathbf{X}}_U^o$ and $\hat{\mathbf{X}}_\gamma^o$ respectively.

7.4 Other considerations

7.4.1 Upper and lower bound proximity analysis

In this section, we study the proximity between the lower and upper bounds. Since $\hat{\mathcal{X}}_{U,k}^o$ and $\hat{\mathcal{X}}_{L,k}^o$ are both positive terms and $\hat{\mathcal{X}}_{U,k}^o > \hat{\mathcal{X}}_{L,k}^o$, we consider the vector

$$\mathbf{B} = (\hat{\boldsymbol{\mathcal{X}}}_U^{o\,[2]} - \hat{\boldsymbol{\mathcal{X}}}_L^{o\,[2]})./\,\text{diag}\{\mathbf{R}_{\mathbf{X}}\} \tag{7.36}$$

as a proximity measure where ./ indicates an element-wise division. Each element B_k of vector $\mathbf{B}$ is therefore a difference of squared values normalized by $\sigma_{X,k}^2 = E\{\mathcal{X}_k^2\}$. From

(7.14), (7.16) and (7.33), we have :

$$\mathbf{B} = \text{diag}\{\mathbf{G}_{\text{MMSE}}\mathbf{R}_{\mathbf{W}}\}./\,\text{diag}\{\mathbf{R}_{\mathbf{X}}\} \tag{7.37}$$

$$= \text{diag}\{\mathbf{R}_{\mathbf{X}}(\mathbf{R}_{\mathbf{X}} + \mathbf{R}_{\mathbf{W}})^{-1}\mathbf{R}_{\mathbf{W}}\}./\,\text{diag}\{\mathbf{R}_{\mathbf{X}}\}. \tag{7.38}$$

Therefore, the second term in (7.33) dictates how tight are the bounds. Interestingly, this term does not depend on $\mathbf{Y}$ (however, in practice, the estimation of $\mathbf{R}_{\mathbf{X}}$ does).

Uncorrelated frequencies

To gain some insight into the behavior of the proximity vector $\mathbf{B}$, let us first consider uncorrelated frequency components. In that case, the k^{th} entry of $\mathbf{B}$ reduces to:

$$B_k = \frac{\sigma^2_{W,k}}{\sigma^2_{X,k} + \sigma^2_{W,k}} = \frac{1}{1 + \text{SNR}_k} \tag{7.39}$$

where $\text{SNR}_k = \sigma^2_{X,k}/\sigma^2_{W,k}$. For a high SNR_k, we have $B_k \to 0$, while for a low SNR_k $B_k \to 1$. Therefore, the bounds will be tighter as the SNR_k gets higher.

Correlated frequencies

We next consider the case of correlated frequency components. $\mathbf{B}$ can be written in a form resembling (7.39):

$$\mathbf{B} = \text{diag}\{\mathbf{R}_{\mathbf{X}}(\mathbf{I}_N + \mathbf{R}_{\mathbf{W}}{}^{-1}\mathbf{R}_{\mathbf{X}})^{-1}\}./\,\text{diag}\{\mathbf{R}_{\mathbf{X}}\} \tag{7.40}$$

Observe that $\mathbf{R}_{\mathbf{X}}(\mathbf{I}_N + \mathbf{R}_{\mathbf{W}}{}^{-1}\mathbf{R}_{\mathbf{X}})^{-1} = \mathbf{R}_{\mathbf{X}}(\mathbf{R}_{\mathbf{W}}{}^{1/2}(\mathbf{R}_{\mathbf{W}}{}^{-1/2}\mathbf{R}_{\mathbf{X}}\mathbf{R}_{\mathbf{W}}{}^{-1/2} + \mathbf{I}_N)\mathbf{R}_{\mathbf{W}}{}^{1/2})^{-1}\mathbf{R}_{\mathbf{W}}$ and let $\mu_{\max} = \mu_N \geq \cdots \geq \mu_1 = \mu_{\min}$ denote the eigenvalues of $\mathbf{R}_{\mathbf{W}}{}^{-1/2}\mathbf{R}_{\mathbf{X}}\mathbf{R}_{\mathbf{W}}{}^{-1/2}$. On the one hand, if $\mu_{\min} >> 1$ (high SNR), then $\mathbf{B} \to \text{diag}\{\mathbf{R}_{\mathbf{W}}\}./\,\text{diag}\{\mathbf{R}_{\mathbf{X}}\}$ while on the other

hand, if $\mu_{\max} << 1$ (low SNR), then $\mathbf{B} \rightarrow \mathbf{1}_{N\times 1}$, where $\mathbf{1}_{N\times 1}$ denotes an N-dimensional column vector of ones. Therefore, again, the bounds will be tighter as the SNR gets higher.

7.4.2 Estimating $\mathbf{R_X}$ and $\mathbf{R_W}$

To compute $\hat{\boldsymbol{\mathcal{X}}}_L^o$ (7.16), $\hat{\boldsymbol{\mathcal{X}}}_U^o$ (7.33) or $\hat{\boldsymbol{\mathcal{X}}}_\gamma^o$ (7.34), one needs an estimation of matrices $\mathbf{R_X}$ and $\mathbf{R_W}$. We shall denote the estimates of $\mathbf{R_X}$, $\mathbf{R_W}$ and $\mathbf{R_Y}$ for the i^{th} frame by $\hat{\mathbf{R}}_{\mathbf{X},i}$, $\hat{\mathbf{R}}_{\mathbf{W},i}$ and $\hat{\mathbf{R}}_{\mathbf{Y},i}$ respectively.

In this work, we use a decision-directed type of approach similar to [4] to estimate $\mathbf{R_X}$. Since $\mathbf{R_X} = E\{\mathbf{X}\mathbf{X}^H\}$ and $\mathbf{R_X} = \mathbf{R_Y} - \mathbf{R_W}$ for uncorrelated $\mathbf{X}$ and $\mathbf{W}$, we have for frame i:

$$\hat{\mathbf{R}}_{\mathbf{X},i} = \tau\hat{\mathbf{X}}_{i-1}\hat{\mathbf{X}}_{i-1}^H + (1-\tau)\rho(\hat{\mathbf{R}}_{\mathbf{Y},i} - \hat{\mathbf{R}}_{\mathbf{W},i}) \tag{7.41}$$

where $\hat{\mathbf{X}}_{i-1}$ is given by (7.10) for frame $i-1$, $0 \leq \tau \leq 1$ is a forgetting factor and $\rho(\cdot)$ is a thresholding function of its matrix argument. In fact, the terms on the diagonal of $\hat{\mathbf{R}}_{\mathbf{X}}$ should be positive, for an $N \times N$ matrix $\mathbf{A}$, we therefore define the lm^{th} element of $\rho(\mathbf{A})$ as:

$$[\rho(\mathbf{A})]_{lm} = \begin{cases} \max([\mathbf{A}]_{lm}, 0) & \text{if } l = m \\ [\mathbf{A}]_{lm} & \text{else} \end{cases} \tag{7.42}$$

The $\max(\cdot,\cdot)$ operator is therefore applied only on the main diagonal of matrix $\hat{\mathbf{R}}_{\mathbf{Y},i} - \hat{\mathbf{R}}_{\mathbf{W},i}$. This approach may result, in practice, in a non-negative definite $\hat{\mathbf{R}}_{\mathbf{X},i}$. A more formal approach, based on eigenvalue decomposition where the eigenvalues are forced to be positive, was also experimented to enforce a non-negative definite constraint. In practice, it was observed that this approach gives similar results to the proposed simplified approach (7.41)-(7.42), however, at a much higher computational cost.

In addition to the estimator $\hat{\mathbf{R}}_{\mathbf{X},i}$ (7.41), we also experimented with a modified structure

for the estimation of $\mathbf{R}_{\mathbf{X},i}$ that takes into account the nature of the current frame, i.e. voiced vs. unvoiced. Indeed, since the correlation due to the harmonics of the fundamental frequency is only present in the voiced parts of speech, it would be appropriate to consider a diagonal $\hat{\mathbf{R}}_{\mathbf{X},i}$ in unvoiced parts and a full (i.e. unconstrained) $\hat{\mathbf{R}}_{\mathbf{X},i}$ in voiced parts. A similar approach was used in [113] where a hard threshold is used to distinguish between voiced and unvoiced speech sections. Here, we propose a soft threshold approach in which the constrained estimator of $\mathbf{R}_{\mathbf{X},i}$, denoted $\hat{\mathbf{R}}_{\mathbf{X},i}^{\delta_i}$, is computed as:

$$\hat{\mathbf{R}}_{\mathbf{X},i}^{\delta_i} = \delta_i \hat{\mathbf{R}}_{\mathbf{X},i} + (1-\delta_i)\,\mathrm{diag}\{\hat{\mathbf{R}}_{\mathbf{X},i}\}. \tag{7.43}$$

where $\hat{\mathbf{R}}_{\mathbf{X},i}$ is given by (7.41) and $0 \leq \delta_i \leq 1$ is a soft threshold parameter accounting for voiced or unvoiced frames. We use the zero-crossing rates (ZCR) in $y_i[n]$ to distinguish between voiced and unvoiced parts since voiced parts are primarily low frequencies and unvoiced parts are primarily high frequencies [50]. A ZCR voiced threshold t_v is used, below which the frame is judged to be voiced and δ_i is set to 1. A ZCR unvoiced threshold $t_u > t_v$ is also used, above which the frame is judged to be unvoiced and δ_i is set to 0. For ZCR between t_u and t_v, intermediate values of δ_i are used. Specifically, the value of δ_i is evaluated as follows:

$$\delta_i = \begin{cases} 1 & \mathrm{ZCR} \leq t_v \\ \frac{t_u - \mathrm{ZCR}}{t_u - t_v} & t_v < \mathrm{ZCR} < t_u \\ 0 & \mathrm{ZCR} \geq t_u. \end{cases} \tag{7.44}$$

The clean speech estimators using $\hat{\mathbf{R}}_{\mathbf{X},i}^{\delta_i}$ (7.43) to estimate $\mathbf{R}_{\mathbf{X},i}$ will be denoted by the additional subscript δ, i.e. $\hat{\mathbf{X}}_{\delta\mathrm{MMSE}}$, $\hat{\mathbf{X}}_{\delta L}^{o}$, $\hat{\mathbf{X}}_{\delta U}^{o}$ and $\hat{\mathbf{X}}_{\delta\gamma}^{o}$, otherwise, the estimator will use $\hat{\mathbf{R}}_{\mathbf{X},i}$ (7.41). We refer to $\hat{\mathbf{R}}_{\mathbf{X},i}^{\delta_i}$ as the soft threshold structured estimator as opposed to the unstructured $\hat{\mathbf{R}}_{\mathbf{X},i}$.

To compute $\hat{\boldsymbol{\mathcal{X}}}_L^o$ (7.16), $\hat{\boldsymbol{\mathcal{X}}}_U^o$ (7.33) or $\hat{\boldsymbol{\mathcal{X}}}_\gamma^o$ (7.34), we also need to estimate $\mathbf{R}_{\mathbf{W},i}$. To do so, we first obtain a time-domain correlation matrix, $\hat{\mathbf{R}}_{\mathbf{w},i} = \mathbf{M}_{\mathbf{w},i}^H \mathbf{M}_{\mathbf{w},i}$ where $\mathbf{M}_{\mathbf{w},i}$ is a matrix whose columns are shifted versions of the time-domain noise data vector (see (8.20) of [118]). Using the $N \times N$ Fourier transform matrix $\mathbf{F}$, we then obtain:

$$\hat{\mathbf{R}}_{\mathbf{W},i} = \mathbf{F}\hat{\mathbf{R}}_{\mathbf{w},i}\mathbf{F}^H. \tag{7.45}$$

Note that while $\hat{\mathbf{R}}_{\mathbf{w},i}$ is a Toeplitz matrix, $\hat{\mathbf{R}}_{\mathbf{W},i}$ will not be Toeplitz in general. $\mathbf{R}_{\mathbf{Y},i}$ is estimated similarly.

7.5 Concluding remarks

In this chapter we considered a multi-dimensional Bayesian STSA estimator for speech enhancement that assumes correlated frequency components. Since its closed-form solution is not readily available, we approached the problem of finding approximations to that estimator from a bounding perspective. We obtained convenient upper and lower bounds and proposed a family of estimators based on these bounds that is parameterized by $0 \leq \gamma \leq 1$.

In Section 8.5, results are presented for wideband PESQ, LLR and informal listening experiments. They demonstrate noticeable advantages, especially at high SNR, of the proposed estimators over existing estimators that consider uncorrelated frequency components, such as the MMSE STSA and Wiener, as well as than an MMSE estimator of the complex STFT coefficients that assumes correlated frequency components. In particular, $\hat{\mathbf{X}}_{\delta\gamma}^o$ offers a good compromise between speech quality and background noise quantity and whiteness.

The work in this chapter was submitted as a journal paper [119] and presented in part in [120].

Chapter 8

Experimental results

In this chapter, we evaluate the estimators developed in this thesis using subjective and objective experimental approaches and compare them with existing relevant estimators. The subjective evaluations include informal listening experiments as well as MOS and MUSHRA measures, while the objective measures include segmental SNR, LLR and PESQ.

In Section 8.1, we explain the methodology used to create the noisy speech signals to which the speech enhancement estimators will be applied. In Section 8.2, we present the MOS and MUSHRA subjective measures as well as the segmental SNR, LLR and PESQ objective measures. Experimental results for the extended β-SA estimator with $\beta > -2$ (Chapter 4), the Wβ-SA estimator with auditory-based parameter values (Chapter 5) and the multi-dimensional STSA estimator allowing for correlated frequency components (Chapter 7) are presented in Section 8.3, 8.4 and 8.5, respectively.

8.1 Creating the noisy speech

The noisy speech signals used in the evaluation of the proposed estimators were constructed by adding noise to clean speech sentences following the additive model presented in Chapter 3. Results using three types of noises from the Noisex database [121], representative of different situations, are presented in this chapter: a so-called white noise, a pink noise and an aircraft cockpit (buccaneer-1) noise. These were available at a sampling frequency of 20kHz and downsampled to either 8kHz or 16kHz, depending on the experiment, using the Matlab function `resample`. The average spectrum magnitudes of those different noises are shown in Fig. 8.1. As can be observed, the white noise has a somewhat flat magnitude, the pink noise magnitude is higher at low frequencies while the cockpit noise is somewhat similar to pink noise but has a significant peak around 2800 Hz. We considered these noises to be quasi-stationary. We chose not to use babble noise and other highly non-stationary noises since the performance of the speech enhancement algorithms would then highly rely on the noise statistics estimation which is not the main topic of this thesis.

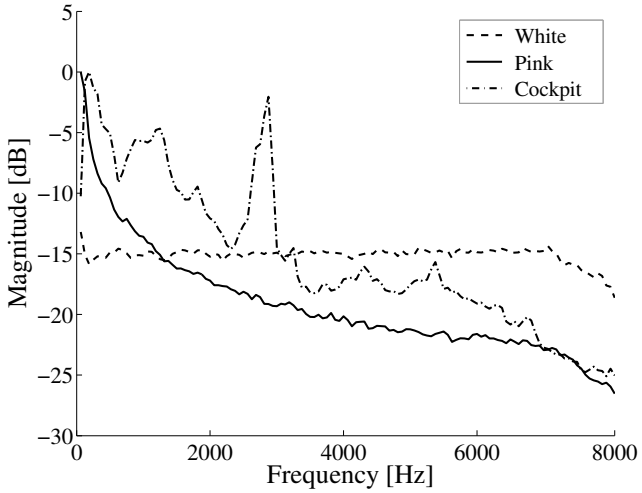

Fig. 8.1 Average noise spectrum magnitudes [dB] versus frequency [Hz] for white, pink and cockpit noises (16 kHz sampling frequency).

Thirty (30) Harvard sentences [122] were used in the experiments (3 males, 3 females, 5 sentences each); see Annexe B for a complete listing of the sentences. These clean speech sentences had an original sampling frequency of 48kHz and, depending on the experiment, were either downsampled at 8kHz or 16kHz, using again the Matlab function `resample`. We added some zero padding at the beginning and end of each sentences to simulate silence periods of 0.75 sec; the total sentences lengths were approximately 4 sec.

The noisy speeches were constructed by adding the scaled samples of the selected noise to the desired clean speech Harvard sentences. The noise level was adjusted to obtain the desired SNR which was evaluated based on the clean speech's "active speech level" according to ITU-T P.56 [123, 124]. In fact, the active speech level takes into account the silence periods of the clean speech and therefore allows the experimenter to adjust the noise power to obtain a more accurate SNR.

8.2 Overview of subjective and objective performance measures

In order to evaluate the performances of speech enhancement estimators, several approaches can be used. In this thesis, we use both subjective and objective approaches. In the former, individuals listen to the enhanced speech and rate them according to different criteria. In objective approaches, the rating is performed by an algorithm that yields a score that is then interpreted to give some indication on the quality of the enhancement.

8.2.1 Subjective measures

We will use different types of subjective evaluations in this chapter, namely, informal listening experiments as well as MOS and MUSHRA measures.

In informal listening experiments, the experimenter, or other participants, listen to the

enhanced speech and give their opinion on different aspects of this audio experience, e.g. the amount of background noise and/or speech distortions. Tests resulting in a quantitative rating of the enhanced speech can also be performed. Those include the family of Absolute Category Ratings (ACR) tests and that of Degradation Category Ratings (DCR) tests [125]. In an ACR test, listeners rate the enhanced audio files using a five level impairment scale. After obtaining individual scores, the mean opinion of all listeners for each audio file is calculated. To achieve reliable results, those tests are performed with a large pool of listeners and under controlled conditions. In DCR tests, listeners hear the reference and the test signals sequentially, and are asked to compare them.

According to [125], major conceptual differences between ACR and DCR tests are that in ACR even an original signal can receive low grade, since listeners compare with their internal model of "clean speech", while DCR tests provide a quality scale of higher resolution, due to comparison of the distorted signal with one or more reference/anchor signals. DCR tests are more common in audio quality assessment [21, 22], while speech coding systems are typically assessed by an ACR test. MOS is a widely used ACR test while MUSHRA is a DCR test.

MOS

MOS (ITU-T P.800 [45]) is a widely used ACR test to evaluate the overall speech quality. It is a five level scale where the listener assigns a value from 1 to 5 to each listened sentence according to Table 8.1 (Overall). The mean of all listeners for each sentence is then evaluated and the final Mean Opinion Score is obtained. As suggested in ITU-T P.835 [126], MOS results can include a separate assessment of the speech distortion where the subject only concentrates on the perceived speech distortion and rates it according to Table 8.1 (Speech). Moreover, it can also include a separate assessment of the background

Table 8.1 MOS scale for speech distortion, background noise and overall appreciation [126].

Scale	Overall	Speech	Background
5	Excellent	Not distorted	Not noticeable
4	Good	Slightly distorted	Slightly noticeable
3	Fair	Somewhat distorted	Noticeable but not intrusive
2	Poor	Fairly distorted	Somewhat intrusive
1	Bad	Very distorted	Very intrusive

noise perception where the subject only concentrate on the perceived background noise and rates it according to Table 8.1 (Background).

MUSHRA

In the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) (ITU-R BS.1534-1 [46]), the subjects are provided with the test utterances plus one reference and one hidden anchor and are asked to rate the different signals on a scale of 0 to 100, 100 being the best score. The rating is performed with the use of slides from a computer user interface (see Fig. 8.2). The hidden anchor is used to provide an indication of how the enhanced files compare to well-known audio quality levels. As the hidden anchor, one may use a signal having an SNR of 5 dB less than the noisy signal to be enhanced, as e.g. in [127]. MUSHRA was originally developed to assess intermediate quality level of coding systems.

8.2.2 Objective measures

Many objective measures are available to assess speech enhancement algorithms [43,77,129]. They are more or less correlated with subjective measures such as MOS. In this thesis, we will use three objective measures, namely the SNR_{seg}, LLR and PESQ measures.

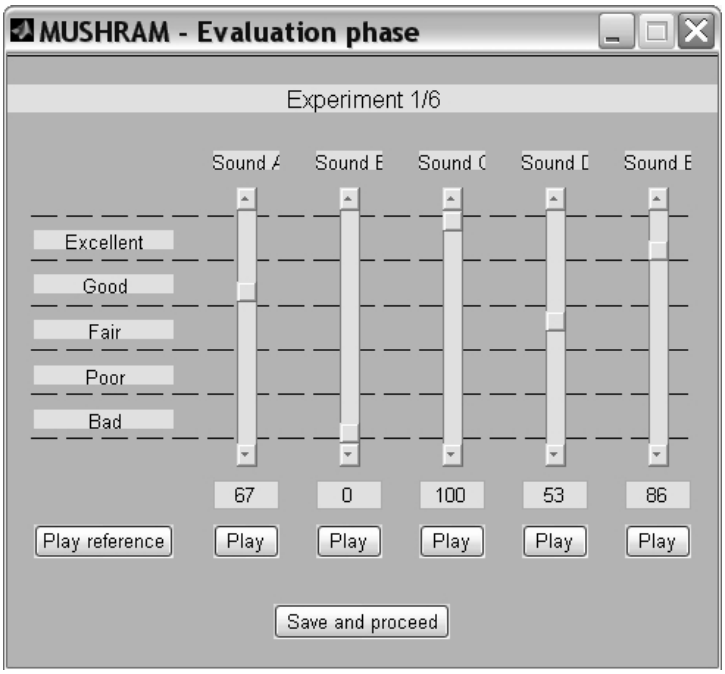

Fig. 8.2 MUSHRA user interface from [128].

$\mathbf{SNR}_{seg}$

The segmental SNR (SNR_{seg}) measure can be expressed as [42]:

$$\text{SNR}_{seg} = \frac{1}{N_f} \sum_{i=0}^{N_f-1} 10 \log_{10} \frac{\|\mathbf{x}_i\|^2}{\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2} \tag{8.1}$$

where $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ are the N-dimensional column vectors comprising of the clean and enhanced speech samples for frame i, respectively, and N_f is the number of frames in the speech signal as before. As proposed in [42], each frame SNR is thresholded by a -20 dB lower bound and a 35 dB higher bound. This is necessary since an SNR above 35 dB does not reflect large perceptual differences while large negative SNRs do not truly reflect the perceptual contribution of the corresponding signal either [130]. A higher value of SNR_{seg} indicates a better performance.

LLR

The Log-Likelihood Ratio (LLR) measure is based on the dissimilarity between all-pole models of the clean speech and the enhanced speech waveforms. It can be expressed for a particular frame i as [43]:

$$\mathrm{LLR}_i = \log\left(\frac{\hat{\mathbf{a}}_i^T \mathbf{R}_{\mathbf{x},i} \hat{\mathbf{a}}_i}{\mathbf{a}_i^T \mathbf{R}_{\mathbf{x},i} \mathbf{a}_i}\right) \tag{8.2}$$

where $\mathbf{a}_i$ is the $(p+1)\times 1$ vector consisting of the linear predictive coding (LPC) coefficients of the original clean speech signal for some order p, $\hat{\mathbf{a}}_i$ is the corresponding LPC coefficient vector of the enhanced speech signal and $\mathbf{R}_{\mathbf{x},i}$ is the autocorrelation matrix of the original clean speech signal in the time domain for frame i, $x_i[n]$. The lm^{th} elements or $\mathbf{R}_{\mathbf{x},i}$ are defined as [43]:

$$r_{x,i}(|l-m|) = \sum_{n=1}^{N-|l-m|} x_i[n]x_i[n+|l-m|], \quad \text{for} \quad |l-m| = 0, 1, \ldots, p. \tag{8.3}$$

The mean LLR for all frames is evaluated from the different LLR_i. To remove unrealistically high distortion levels, the frames with LLR greater than five times the standard deviation of all LLR_i are ignored in the averaging process [130] (this corresponded typically to less than 1% of the frames). A lower LLR score indicates a better performance (e.g. a score of 0 is obtained for identical clean and enhanced speech). The LLR was computed in this thesis using algorithms made available by the Robust Speech Processing Laboratory of Duke University[1].

[1]These files were originally obtained from `http://cslr.colorado.edu/rspl/rspl_software.html` in July 2006 but are not available anymore.

PESQ

PESQ [44] aims at predicting the perceived quality of a test sentence if it was evaluated in a subjective listening test such as MOS and is designed to yield a score from 1 to 4.5. It was originally intended for automatically measuring the quality of narrow band telephone signals, but its realm of applications has been extended since then. To predict the subjective score, the PESQ algorithm uses the clean speech signal, i.e. without any noise component, along with the enhanced speech signal. The two signals are processed by a perceptual model and the final score is derived using the perceptual representation of the clean and enhanced speech signal. While the PESQ measure has not been approved to assess speech enhancement algorithms, it has lately been widely used to do so [21, 29, 127].

The original PESQ measure was developed for 8 kHz sampled speech. In order to extend the application of PESQ to systems such as wideband telephony and speech codecs, a wideband extension to the PESQ measure has also been proposed in ITU-T P.862.2 [131].

A study of the correlation between MOS and some objective measures was presented in [132]. Results showed that the SNR_{seg} measure is poorly correlated with the overall quality of the enhanced speech (see Table 8.2) and signal distortion but better correlated with the background noise. Moreover, the LLR measure was found to be correlated with both overall quality and speech distortion and a little less correlated with the noise reduction. Finally, the PESQ measure was found to be well correlated with all three aspects. Therefore, a fairly small difference in terms of PESQ (e.g. ± 0.05) or LLR (e.g. ± 0.05) between two enhanced speech should most likely correspond to a significant perceivable difference as if it would be measured by MOS.

Table 8.2 Estimated correlation coefficient of SNR_{seg}, LLR and PESQ objective measures with overall quality, signal distortion, and background noise [133].

	Overall Quality	Signal Distortion	Background Noise
SNR_{seg}	0.36	0.22	0.56
LLR	0.85	0.88	0.51
PESQ	0.89	0.81	0.76

8.3 Evaluation of the extended β-SA estimator

In this section, we present evaluation results for the β-SA estimator with negative values of β as discussed in Chapter 4. We present comparative results for three estimators: MMSE STSA (or β-SA with $\beta = 1$), LSA (or β-SA with $\beta \to 0$) and β-SA with $\beta = -1$. It was found through informal listening experiments that the value of $\beta = -1$ offered a good compromise between noise reduction and speech distortion; however, serious speech distortions were introduced when β became smaller than -1.5. This motivated the choice of $\beta = -1$ for comparisons.

8.3.1 Methodology

In our experiment, the sampling rate was set to 8kHz. As explained in Section 8.1, thirty Harvard sentences [122] (3 men and 3 women each speaking 5 sentences) were used as the clean speech. These were corrupted by additive noise, i.e. white, pink and cockpit noises [121], scaled to obtain the desired SNR (i.e. 0 dB, 5 dB and 10 dB). For each combination of SNR and noise type, the noisy sentences were processed with the MMSE STSA (β-SA with $\beta = 1$), LSA (β-SA with $\beta \to 0$) and β-SA with $\beta = -1$ and the enhanced files were evaluated using the PESQ objective measure and the MOS subjective measure.

In our implementation of the STSA estimators, the frame duration was set to $N = 256$ samples (32ms). The overlap-add method with a 50% overlap between adjacent frames was used to synthesize the enhanced speech. All algorithms in this subsection, and also in Subsection 8.4, used the *decision-directed* approach for the estimation of ξ_k (3.54). In fact, its the most widely used method to estimate ξ_k encountered in the literature and, more specifically, the one used in [29]. The voice activity detector (VAD) proposed in [90] was used in the evaluation of the noise variance.

Referring to the β-SA gain in (4.7), the confluent hypergeometric function, $M(a, b; z)$, was implemented using the *chgm* function from [134] which was converted to the Matlab language. For values of $\upsilon_k > 700$, we implemented an approximation to the confluent hypergeometric function as given by (13.1.5) from [104]. This was necessary to avoid *NaN* and *Inf* output from the *chgm* function. This approximation is also used in the next sections of this thesis.

8.3.2 Results and discussion

Table 8.3 presents the PESQ results for the three types of noises at SNRs of 0 dB, 5 dB and 10 dB. As can be observed, the β-SA with $\beta = -1$ slightly outperforms MMSE STSA and LSA except for pink noise at a 10 dB SNR, where LSA shows a slight advantage, and for cockpit noise at 10 dB, where both approaches show an equivalent performance. These results indicate that the β-SA estimator with $\beta = -1$ is advantageous at lower SNR.

Since the PESQ values of the LSA and β-SA (β = -1) are close in Table 8.3, a subjective assessment needs to be performed in order to identify if the differences are significant. In order to support the results obtained with PESQ, we performed informal MOS subjective listening tests on 6 subjects using a subset of 4 sentences from the initial 30, each spoken by a different individual (2 men, 2 women). Therefore, the average for each final MOS

Table 8.3 PESQ results for MMSE STSA, LSA and β-SA ($\beta = -1$) estimators for white, pink and cockpit noises at several SNRs (0 dB, 5 dB and 10 dB).

		Noisy speech	**MMSE STSA**	**LSA**	**β-SA ($\beta = -1$)**
white	0 dB	1.29	1.39	1.44	**1.47**
	5 dB	1.37	1.60	1.70	**1.72**
	10 dB	1.58	1.83	1.95	**1.96**
pink	0 dB	1.35	1.54	1.64	**1.68**
	5 dB	1.50	1.78	1.91	**1.94**
	10 dB	1.79	2.00	**2.14**	2.13
cockpit	0 dB	1.29	1.46	1.53	**1.57**
	5 dB	1.44	1.67	1.78	**1.81**
	10 dB	1.67	1.91	**2.03**	**2.03**

Table 8.4 Informal MOS results for MMSE STSA, LSA and β-SA ($\beta = -1$) estimators (SNR = 0dB).

		Noisy speech	**MMSE STSA**	**LSA**	**β-SA ($\beta = -1$)**
white	Speech	3.9	2.4	**2.9**	2.8
	Background	1.2	2.2	2.5	**2.9**
	Overall	1.7	2.1	2.5	**2.7**
cockpit	Speech	3.7	2.8	**3.1**	2.8
	Background	1.2	2.4	2.8	**2.9**
	Overall	1.8	2.4	**2.8**	2.5

score is made over 24 scores. As suggested by ITU-T P.835 [126], MOS tests included an assessment of the speech distortion (5 = Not distorted, 1 = Very distorted), background noise (5 = Not noticeable, 1 = Very intrusive) and overall speech quality (5 = Excellent, 1 = Bad). Tests were performed in an isolated acoustic room using high quality *beyerdynamic DT880* headphones.

Table 8.4 presents the informal MOS test results. When comparing the LSA and β-

SA with $\beta = -1$, we see that the latter demonstrated more speech distortion but also more noise reduction than the former for both noises, as expected from Subsection 4.2.2. However, the overall perception was not the same for both noises. In fact, β-SA with $\beta = -1$ was thought to be better than LSA for white noise but the inverse was found for cockpit noise. Also, based on Fig. 4.2, MMSE STSA should have yielded less speech distortion (i.e. higher score for *Speech* in Table 8.4) than the other two estimators, however, this is not what we have observed. This could be due to the fact that, when a frame overlap of 50% is used, a perceivable echo is present in the MMSE STSA enhanced signal which is quite less perceivalble in LSA and β-SA and may not have been well taken into account by (4.9).

In summary, we showed that, when setting $\beta = -1$ in the β-SA estimator, the latter achieves better results in terms of PESQ than the MMSE STSA and LSA estimators for low SNR values. Also, the overall MOS appreciation for the β-SA estimator with $\beta = -1$ is found to be better for white noise but inferior than the LSA estimator for the cockpit noise.

8.4 Evaluation of the Wβ-SA with auditory-based parameter values

In this section, we present the experimental results for the Wβ-SA estimator with the proposed auditory-based choice of β and α values as developed in Chapter 5. We will compare it to the MMSE STSA, LSA and WE estimators using both objective and subjective measures.

8.4.1 Methodology

The same three types of noises as before are used in these experiments: a white noise and two colored noises, that is a pink noise and an aircraft cockpit noise [121]. Other noise types were considered during our experimentation and lead to the same conclusions as the ones drawn below. The number of noisy sentences used respectively in the objective and subjective evaluations will be specified in the corresponding subsections below.

All speech signals were sampled at 16 kHz in this section and a raised-cosine window [67] was used (512 samples, 32ms) in the STSA computation. A 75% overlap was used in the overlap-add synthesis method as in [4] to limit the reverberation effect in the MMSE STSA estimator noted in the previous section. As in the previous subsection, all algorithms used the *decision-directed* approach [4] for the estimation of the *a priori* SNR ξ_k (3.54), its the most widely used method to estimate ξ_k encountered in the literature and, more specifically, the one used in [6, 29]. The VAD algorithm proposed in [90] was used in the evaluation of the noise spectral amplitude variance.

The noisy speech files were processed with the proposed Wβ-SA estimator, as well as with the MMSE STSA, LSA and WE estimators. The value of p in the WE estimator was set to $p = -1$ as proposed in [6]. We do not consider explicitly the β-SA estimator for comparison, however, we note that the case $\alpha = 0$ in the Wβ-SA corresponds to the β-SA estimator.

As mentioned in Subsection 3.1.2, the STFT framework for speech enhancement produces some time domain aliasing. To limit the latter, we can append zeros at the end of the window function $h_a[n]$ in (3.8). When doing so with the proposed Wβ-SA estimator, we noticed that perceptible discontinuities between two adjacent frames could occur depending on the gain difference between the two frames. To limit these discontinuities, instead

of appending N zeros at the end of each frame, we appended $N/2$ zeros at the beginning and $N/2$ zeros at the end. This amounts to shifting the original time-domain zero-padded signal by $N/2$ samples and is similar to the approach in [25] (also known as a causality delay). It had the effect of limiting greatly the discontinuities between adjacent frames.

Results in this section are presented for the MUSHRA subjective measure as well as for the SNR_{seg}, LLR and wideband PESQ objective measures. Since the SNR_{seg} is poorly correlated with the overall quality of the enhanced speech (see Table 8.2), we use it here mainly for discussion regarding the noise reduction of the estimators.

8.4.2 Objective results

Tables 8.5, 8.6 and 8.7 present the SNR_{seg} results for white, pink and cockpit noises, respectively, at an SNR of 0 dB. All SNR_{seg}, LLR and PESQ results are averages obtained from 30 Harvard sentences (3 males, 3 females, 5 sentences each) [122]. The columns and lines of the tables are structured in somewhat decreasing β and increasing α order, where $0.2 \leq \beta_k \leq 1$ refers to the auditory-based selection in (5.11) and $0.5 \leq \alpha_k \leq 0.9$ refers to the choice in (5.12).

Table 8.5 SNR_{seg} for several β and α values (white noise, 0 dB).

	$\beta = 1$	$\beta = \beta_k$	$\beta = 1/3$	$\beta \to 0$
$\alpha = 0$	-0.43 (MMSE STSA)	0.56	0.70	1.12 (LSA)
$\alpha = 0.5$	2.06 (WE $p = -1$)	2.41	2.39	2.57
$\alpha = \alpha_k$	2.46	2.73	2.78	**2.97**

As reported in [6] for the WE estimator and in [29] for the β-SA estimator, we can observe that the SNR_{seg} generally increases for a decreasing β and an increasing α. This result is easily explained since for a decreasing β and an increasing α, the gain function

Table 8.6 SNR_{seg} for several β and α values (pink noise, 0 dB).

	$\beta = 1$	$\beta = \beta_k$	$\beta = 1/3$	$\beta \to 0$
$\alpha = 0$	-1.04 (MMSE STSA)	-0.65	-0.26	0.01 (LSA)
$\alpha = 0.5$	0.52 (WE $p = -1$)	0.62	0.68	0.78
$\alpha = \alpha_k$	0.56	0.63	0.73	**0.84**

Table 8.7 SNR_{seg} for several β and α values (cockpit noise, 0 dB).

	$\beta = 1$	$\beta = \beta_k$	$\beta = 1/3$	$\beta \to 0$
$\alpha = 0$	-1.41 (MMSE STSA)	-0.98	-0.74	-0.50 (LSA)
$\alpha = 0.5$	-0.04 (WE $p = -1$)	0.11	0.11	0.20
$\alpha = \alpha_k$	0.01	0.14	0.16	**0.26**

of the estimator, G_k, decreases (see Fig. 5.1 and Fig. 5.2). In turn, this produces more noise reduction and, as we mentioned previously, the SNR_{seg} is better correlated with noise reduction. The best result is therefore obtained for the smallest β (i.e. $\beta \to 0$) and biggest α (i.e. $\alpha = \alpha_k$).

We present LLR results in Tables 8.8, 8.9 and 8.10 for white, pink and cockpit noises, respectively, at an SNR of 0 dB. For the white noise case, the best results (smallest LLR) were obtained for $\alpha = \alpha_k$. For the colored noises, the best results were obtained for $\beta = \beta_k$, $\alpha = 0.5$. Setting $\alpha = \alpha_k$ reduces greatly the noise at high frequency since it decreases the gain, but it simultaneously introduces some speech distortions, especially when combined to smaller β values. Those high frequency speech distortions were less perceptible in white noise which has a high frequency content. However, for the colored noises used here, which have a small high frequency content, the speech distortions became more perceptible.

We next compare the wideband PESQ results of the proposed Wβ-SA estimator with auditory-based parameters α_k and β_k, with those of the MMSE STSA, LSA and WE

Table 8.8 LLR for several β and α values (white noise, 0 dB).

	$\beta = 1$	$\beta = \beta_k$	$\beta = 1/3$	$\beta \to 0$
$\alpha = 0$	1.96 (MMSE STSA)	1.76	1.85	1.82 (LSA)
$\alpha = 0.5$	1.75 (WE $p = -1$)	1.62	1.73	1.71
$\alpha = \alpha_k$	**1.37**	1.41	**1.37**	1.61

Table 8.9 LLR for several β and α values (pink noise, 0 dB).

	$\beta = 1$	$\beta = \beta_k$	$\beta = 1/3$	$\beta \to 0$
$\alpha = 0$	1.38 (MMSE STSA)	1.25	1.33	1.30 (LSA)
$\alpha = 0.5$	1.28 (WE $p = -1$)	**1.20**	1.27	1.27
$\alpha = \alpha_k$	1.22	1.53	1.40	1.71

Table 8.10 LLR for several β and α values (cockpit noise, 0 dB).

	$\beta = 1$	$\beta = \beta_k$	$\beta = 1/3$	$\beta \to 0$
$\alpha = 0$	1.43 (MMSE STSA)	1.31	1.39	1.38 (LSA)
$\alpha = 0.5$	1.37 (WE $p = -1$)	**1.30**	1.38	1.38
$\alpha = \alpha_k$	1.34	1.60	1.50	1.75

($p = -1$) estimators at noisy speech SNR's between -5 and 5 dB. Fig. 8.3 (a), (b) and (c) show the wideband PESQ improvements[2] over the noisy speech signal wideband PESQ values for the given estimators as a function of SNR, for white noise, pink noise and aircraft cockpit noise respectively. The noisy speech wideband PESQ values were 0.94 at -5 dB and 1.52 at 5 dB (evaluated as averages of all three noise types). For clarity purposes, only the $\beta = \beta_k$, $\alpha = \alpha_k$ case is plotted.

The Wβ-SA estimator with auditory-based parameters β_k, α_k was found to be consistently better than the MMSE STSA, LSA and WE ($p = -1$) estimators in terms of PESQ.

[2] i.e. the wideband PESQ of the enhanced speech minus the wideband PESQ of the noisy speech.

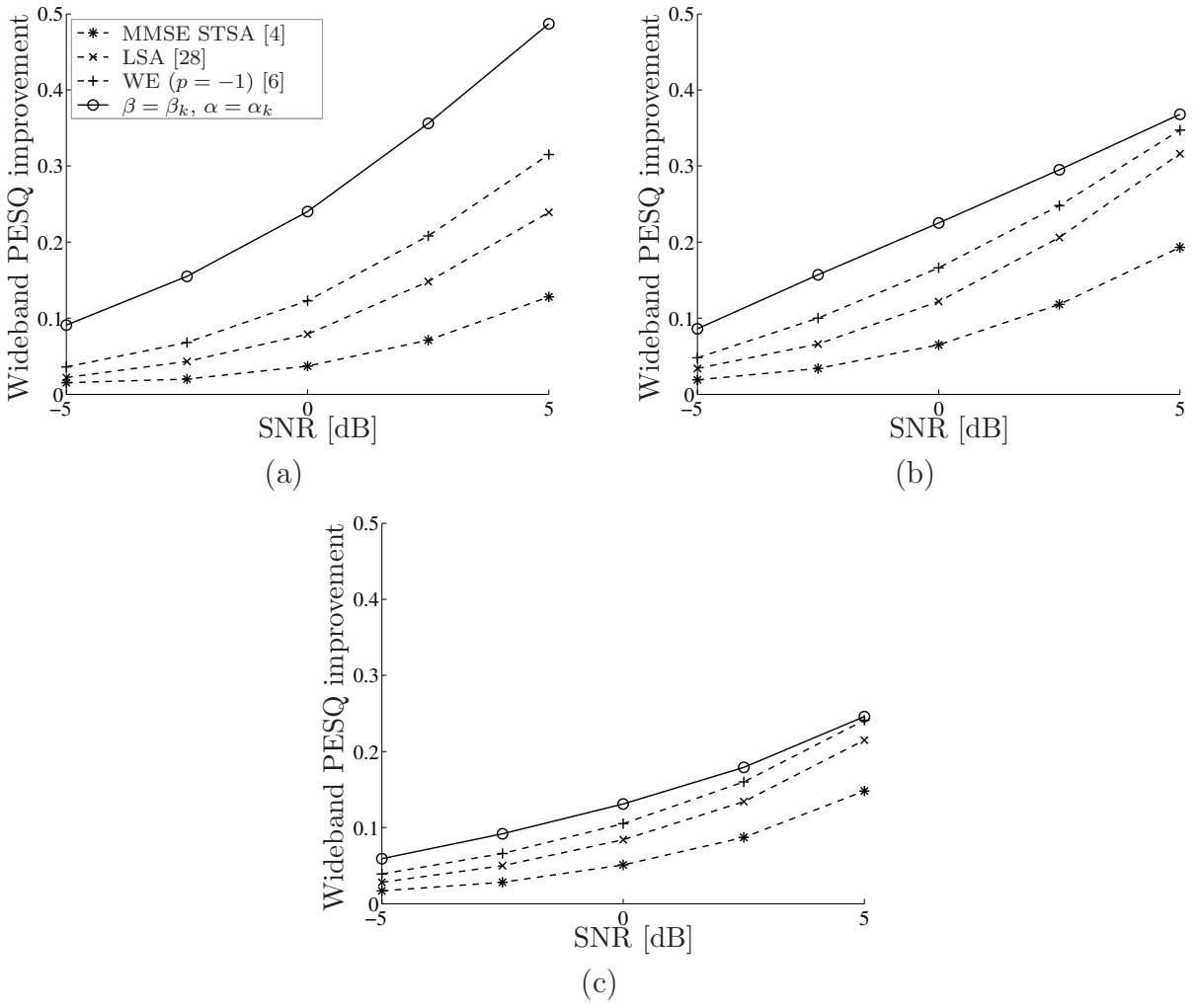

Fig. 8.3 Wideband PESQ improvement over noisy signal versus SNR for (a) white noise, (b) pink noise and (c) aircraft cockpit noise.

While the results are not presented here, the Wβ-SA estimators with $\beta = \beta_k$, $\alpha = 0.5$ and $\beta = 1$, $\alpha = \alpha_k$ where found to be better overall than the WE ($p = -1$) and LSA estimators but not as good as the Wβ-SA with $\beta = \beta_k$, $\alpha = \alpha_k$. The Wβ-SA estimator with $\beta \to 0$, $\alpha = \alpha_k$ and $\beta = 1/3$, $\alpha = \alpha_k$ performed better than LSA and WE ($p = -1$) at an SNR of -5 dB but worst at higher SNRs. In fact, while the case $\beta \to 0$, $\alpha = \alpha_k$ had the highest SNR_{seg} score, it introduces significant speech distortion (as identified by

the LLR results) and shows a poor wideband PESQ value, in particular at higher SNRs. The best compromise is therefore obtained with the auditory-based parameter values, i.e. $\beta = \beta_k$, $\alpha = \alpha_k$.

While the results for male and female spoken utterances are grouped together in the previous tables and figures, an analysis was performed where the results were separated according to the speaker's gender. Results in terms of LLR where similar for both male and female while SNR_{seg} results from the sentences spoken by males were approximately 1 dB inferior to the ones spoken by females; however, the conclusions did not change when comparing the different estimators in each gender group. Wideband PESQ values were found to be slightly inferior for females when compared to males for all estimators. Again, the same ordering of the different estimators was obtained in each group. The only exception was for the cockpit noise and male utterances where the LSA estimator was found to be better than WE for all SNRs and also better than Wβ-SA ($\beta = \beta_k$, $\alpha = \alpha_k$) for an SNR of 5 dB.

8.4.3 Subjective results

As a subjective measure, we used a test setup similar to the MUSHRA test as implemented in [128]. A total of 8 listeners (7 males, 1 female aged in the mid 20's to low 30's with a background in either speech processing or telecommunications) participated in the test of which half where judged to be experienced listeners. The listeners were allowed to listen to each sentence several times and always had access to the clean signal reference. A subset of two sentences (one male speaker, one female speaker) were chosen randomly[3] from the sentences used previously for the objective evaluation. These two sentences were corrupted

[3]A numerical value from 1 to 15 was assigned to each of the sentences in both the female and male uttered sets and the Matlab function `15*rand` was used to select the chosen sentence in each set.

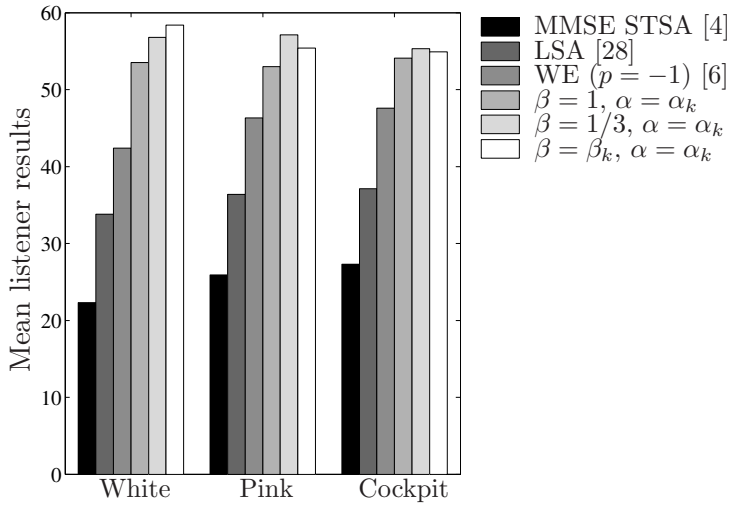

Fig. 8.4 Comparative subjective results for white, pink and cockpit noises (0 dB).

by the same three noise types as before and enhanced using several estimators, the same two sentences were used for all subjects. Tests were performed in an isolated acoustic room using *beyerdynamic* DT880 headphones. The average duration of a test was approximately 30 minutes per subject.

Fig. 8.4 presents the comparative subjective results for the MMSE STSA, LSA and WE estimators along with those of the Wβ-SA estimator with proposed values $\beta = 1$, $\alpha = \alpha_k$, $\beta = 1/3$, $\alpha = \alpha_k$ and $\beta = \beta_k$, $\alpha = \alpha_k$. As can be observed, the sentences enhanced using the Wβ-SA estimator were rated higher than those enhanced by the other estimators for all noise types. Two-tailed paired t-tests [135] revealed the advantage of the Wβ-SA estimator with the proposed values ($\beta = 1$, $\alpha = \alpha_k$; $\beta = 1/3$, $\alpha = \alpha_k$; $\beta = \beta_k$, $\alpha = \alpha_k$) over the WE ($p = -1$) to be statistically significant for all three noise types within a 95% confidence interval.

We observe that listeners in the previous experiment preferred an enhanced speech having more high frequency noise reduction than one having less high frequency speech

distortion. It was observed in [133] that listeners seem to be more sensitive to speech distortion than noise reduction when participating in a subjective evaluation of enhanced speech. This conclusion was based on experiments with sampled speech at 8 kHz whereas we used a 16 kHz sampling rate in our experiments. Therefore, the conclusions of [133] only applies to the lower frequency portion of the spectrum considered in our work. Based on our experimental work with 16 kHz sampling and the result in [133], it would seem that the high frequency speech distortion is less important in subjective evaluations than the low frequency speech distortion.

Additional subjective tests (not shown here), using a smaller subset of the previous subjects, were also performed for an SNR of 5 dB. The Wβ-SA algorithms still received higher scores than all the other algorithms. However, while a substantial advantage of the Wβ-SA estimators was still found over LSA, the difference between the Wβ-SA estimators and the WE ($p = -1$) estimator was found to be narrower than for the 0 dB case. Moreover, an analysis where the results were grouped according to the speaker's gender was also performed for the subjective results. No differences were observed in the comparative results except that the ranking of the three Wβ-SA estimators (i.e. $\beta = 1$, $\alpha = \alpha_k$; $\beta = 1/3$, $\alpha = \alpha_k$ and $\beta = \beta_k$, $\alpha = \alpha_k$) were interchanged for the cockpit noise and male spoken utterances.

8.4.4 Discussion

The human ear is more sensitive between 3 kHz and 4 kHz, as can be observed from an equal loudness curve [54], and will therefore perceive weaker sounds in that frequency band. Therefore, it would seem advantageous to improve the estimation of those weaker sounds in the frequency band between 3 and 4 kHz. Additional experiments where conducted where we locally increased the value of α for those frequencies, therefore giving more importance

to weaker sounds. We compared this approach with the approach using the proposed α_k values. A slight improvement was observed in terms of wideband PESQ for the white noise as well as in terms of LLR for the colored noise cases; all SNR_{seg} values as well as the other wideband PESQ and LLR values showed no significant differences. Moreover, informal listening experiments revealed marginal differences between the two approaches.

We chose the Wβ-SA estimator parameters based on characteristics of the human auditory system. It turn out that, both the approaches using β_k and α_k produce a decrease in the gain G_k at high frequencies compared to lower frequencies (as can be observed from the gains in Fig. 5.1 and Fig. 5.2 with the values of β_k and α_k as in Fig. 5.3 and Fig. 5.4 respectively). This decrease in G_k generates more noise reduction at high frequencies but has the simultaneous effect of producing more speech distortions. The speech distortions are however minimized at low frequencies, where the main speech energy is located, by keeping β high and α low therefore producing a higher gain. The proposed β_k and α_k values will therefore be more advantageous when the noise has high frequency content, such as white noise, in which case more noise will be removed while speech distortions will be less perceptible. This explains why the proposed algorithms obtained the best performance in white noise.

Moreover, the distortions of the high frequency contents of speech, such as fricatives, will be less perceptible in heavy noise (i.e. low SNRs) but they could become more perceptible in regions or sentences where the noise is weak. This could explain why the estimators are more advantageous at smaller SNRs, as observed. It is important to note, however, that the gain is mostly decreased for low instantaneous SNRs. In fact, for high instantaneous SNRs, all the estimator gains tend toward the Wiener gain therefore reducing the speech distortions. For low instantaneous SNRs, the heavy noise will mask the speech signal; since these cannot be restored, the estimator will apply a small gain which will remove much of

the noise.

In summary, improvements over existing Bayesian estimators such as the MMSE STSA, LSA and WE estimators were reported, both in terms of objective (SNR_{seg}, LLR and wideband PESQ) and subjective measures particularly for noise having high frequency content and at low SNRs. Choosing $\beta = \beta_k$ and $\alpha = \alpha_k$ was found to yield good overall results.

8.5 Evaluation of the multi-dimensional estimators for correlated frequency components

In this section, we present experimental results for the multi-dimensional estimators that allow for correlated frequency components developed in Chapter 7. They are compared with conventional Wiener and MMSE STSA, i.e. that both consider uncorrelated frequency components, as well as with an MMSE estimator of the complex STFT coefficients that assumes correlated frequency components. We choose to compare the proposed multi-dimensional estimators with other MMSE estimators[4] and not with ones that use more elaborate cost functions such as the proposed Wβ-SA. In fact, more elaborate cost functions could also be implemented in the multi-dimensional framework and the corresponding estimators could then be compared with their equivalent uncorrelated counterparts.

8.5.1 Methodology

As for the previous experimental results, we use three types of noise from the Noisex database [121]: a white noise and two colored noises, that is a pink noise and an aircraft cockpit noise (buccaneer-1). Noisy speech signals were created according to ITU-T standard

[4]As mentioned in Section 3.3, the Wiener estimator is the MMSE estimator of the STFT assuming uncorrelated frequency components.

P.56 [123]. Thirty noisy sentences (15 from 3 different female speakers and 15 from 3 different male speakers) were used in the evaluations. All speech signals were sampled at 16 kHz and a raised-cosine window [67] was used ($N = 512$ samples, 32ms) in the STSA computation. A 75% overlap was used in the overlap-add synthesis method as in [4]. For the value of simplicity, $\mathbf{R_W}$ was estimated from the first five frames of the noisy speech signal which did not contain any speech signal and its value was kept constant for all subsequent frames.

The value of $\gamma = 0.5$ will be considered in the $\hat{\mathbf{X}}^o_\gamma$ and $\hat{\mathbf{X}}^o_{\delta\gamma}$ estimators. Also, we identified through experimentation the following ZCR thresholds to be used in (7.44): $t_v = 3500$ crossings/sec and $t_u = 6000$ crossings/sec. Since ZCR are affected when the SNR is very low, $\hat{\mathbf{R}}^\delta_{\mathbf{X},i}$ (7.43) was only used if the power of the current frame was 1.5 times the estimated power of the noise, otherwise we used $\hat{\mathbf{R}}_{\mathbf{X},i}$ (7.41). We also set the forgetting factor in (7.41) to $\tau = 0.98$.

Results in this section are presented for informal listening experiments as well as for the LLR and wideband PESQ objective measures.

8.5.2 Informal listening experiments

Informal listening experiments were first conducted to evaluate the qualitative merits of the proposed estimators. The following observations were made:

- The overall difference between $\hat{\mathbf{X}}_{\text{MMSE}}$ and $\hat{\mathbf{X}}^o_L$ was found to be quite small except for some little background clicks that were sometimes present in $\hat{\mathbf{X}}_{\text{MMSE}}$ but not in $\hat{\mathbf{X}}^o_L$. This similarity was expected since only the phase differs between the two estimators (respectively (7.13) and (7.10, 7.16)).

- The traditional Wiener and MMSE STSA estimators exhibit some small perceptible

reverberations that are not present in the novel multi-dimensional estimators considering full correlation matrices i.e. $\hat{\mathbf{X}}_{\text{MMSE}}$, $\hat{\mathbf{X}}_L^o$, $\hat{\mathbf{X}}_U^o$ and $\hat{\mathbf{X}}_\gamma^o$.

- The processed speech in $\hat{\mathbf{X}}_{\text{MMSE}}$, $\hat{\mathbf{X}}_L^o$, $\hat{\mathbf{X}}_U^o$ and $\hat{\mathbf{X}}_\gamma^o$ sounds a little bit more muffled than the one obtained by Wiener or MMSE STSA. By allowing a better model for the unvoiced speech parts, the estimators $\hat{\mathbf{X}}_{\delta\text{MMSE}}$, $\hat{\mathbf{X}}_{\delta L}^o$, $\hat{\mathbf{X}}_{\delta U}^o$ and $\hat{\mathbf{X}}_{\delta\gamma}^o$ better preserve the fricatives and have a less muffled speech.

- The background noise in MMSE STSA, $\hat{\mathbf{X}}_U^o$, $\hat{\mathbf{X}}_\gamma^o$, $\hat{\mathbf{X}}_{\delta U}^o$ and $\hat{\mathbf{X}}_{\delta\gamma}^o$ is whiter than the one of the other estimators which is more musical. Moreover, $\hat{\mathbf{X}}_U^o$ has always more background noise than $\hat{\mathbf{X}}_L^o$, however it is whiter, and MMSE STSA has much more background noise than all other estimators.

- The best estimator overall is found to be the $\hat{\mathbf{X}}_{\delta\gamma}^o$ estimator. In fact, it has whiter background noise than Wiener's, less background noise than MMSE STSA and less speech distortions than the unconstrained full matrix equivalent $\hat{\mathbf{X}}_\gamma^o$.

8.5.3 Objective results

Table 8.11 presents wideband PESQ results for the MMSE STSA, Wiener, $\hat{\mathbf{X}}_{\text{MMSE}}$ and the proposed estimators. As can be observed, the best results for all cases are obtained by one of the proposed estimators. The algorithms that used the soft threshold structured estimator $\hat{\mathbf{R}}_{\mathbf{X},i}^{\delta_i}$ for the clean speech correlation matrix estimation (i.e. $\hat{\mathbf{X}}_{\delta\text{MMSE}}$, $\hat{\mathbf{X}}_{\delta L}^o$, $\hat{\mathbf{X}}_{\delta U}^o$ and $\hat{\mathbf{X}}_{\delta\gamma}^o$) gave better results than the ones using the unstructured $\hat{\mathbf{R}}_{\mathbf{X},i}$ (i.e. $\hat{\mathbf{X}}_{\text{MMSE}}$, $\hat{\mathbf{X}}_L^o$, $\hat{\mathbf{X}}_U^o$ and $\hat{\mathbf{X}}_\gamma^o$) for white noise while they were found more or less equivalent for colored noises. $\hat{\mathbf{X}}_{\delta\gamma}^o$ mostly gave better results than $\hat{\mathbf{X}}_{\delta L}^o$ and $\hat{\mathbf{X}}_{\delta U}^o$ while the advantage of $\hat{\mathbf{X}}_\gamma^o$ over $\hat{\mathbf{X}}_L^o$ and $\hat{\mathbf{X}}_U^o$ was case dependent.

Table 8.11 Wideband PESQ results for white, pink and cockpit noises at several SNRs (10, 15 and 20 dB).

		MMSE STSA [4]	Wiener (7.17)	$\hat{\mathbf{X}}_{\text{MMSE}}$ (7.13)	$\hat{\mathbf{X}}_L^o$	$\hat{\mathbf{X}}_U^o$	$\hat{\mathbf{X}}_\gamma^o$	$\hat{\mathbf{X}}_{\delta\text{MMSE}}$	$\hat{\mathbf{X}}_{\delta L}^o$	$\hat{\mathbf{X}}_{\delta U}^o$	$\hat{\mathbf{X}}_{\delta\gamma}^o$
White	10 dB	1.35	1.53	1.57	1.57	1.46	1.52	**1.61**	**1.61**	1.52	1.59
	15 dB	1.70	1.90	1.94	1.98	1.93	1.98	1.98	2.01	2.04	**2.11**
	20 dB	2.25	2.45	2.39	2.44	2.53	2.52	2.48	2.51	2.62	**2.65**
Pink	10 dB	1.47	1.58	1.70	1.74	1.70	1.74	1.71	1.75	1.71	**1.77**
	15 dB	1.90	1.95	2.05	2.10	2.21	2.20	2.06	2.11	2.19	**2.23**
	20 dB	2.48	2.48	2.49	2.53	**2.72**	2.66	2.55	2.58	2.70	**2.72**
Cockpit	10 dB	1.35	1.38	1.47	1.50	1.52	**1.53**	1.47	1.50	1.50	**1.53**
	15 dB	1.69	1.65	1.77	1.82	**1.94**	1.92	1.76	1.81	1.90	1.91
	20 dB	2.20	2.11	2.19	2.24	**2.43**	2.38	2.22	2.26	2.39	2.39

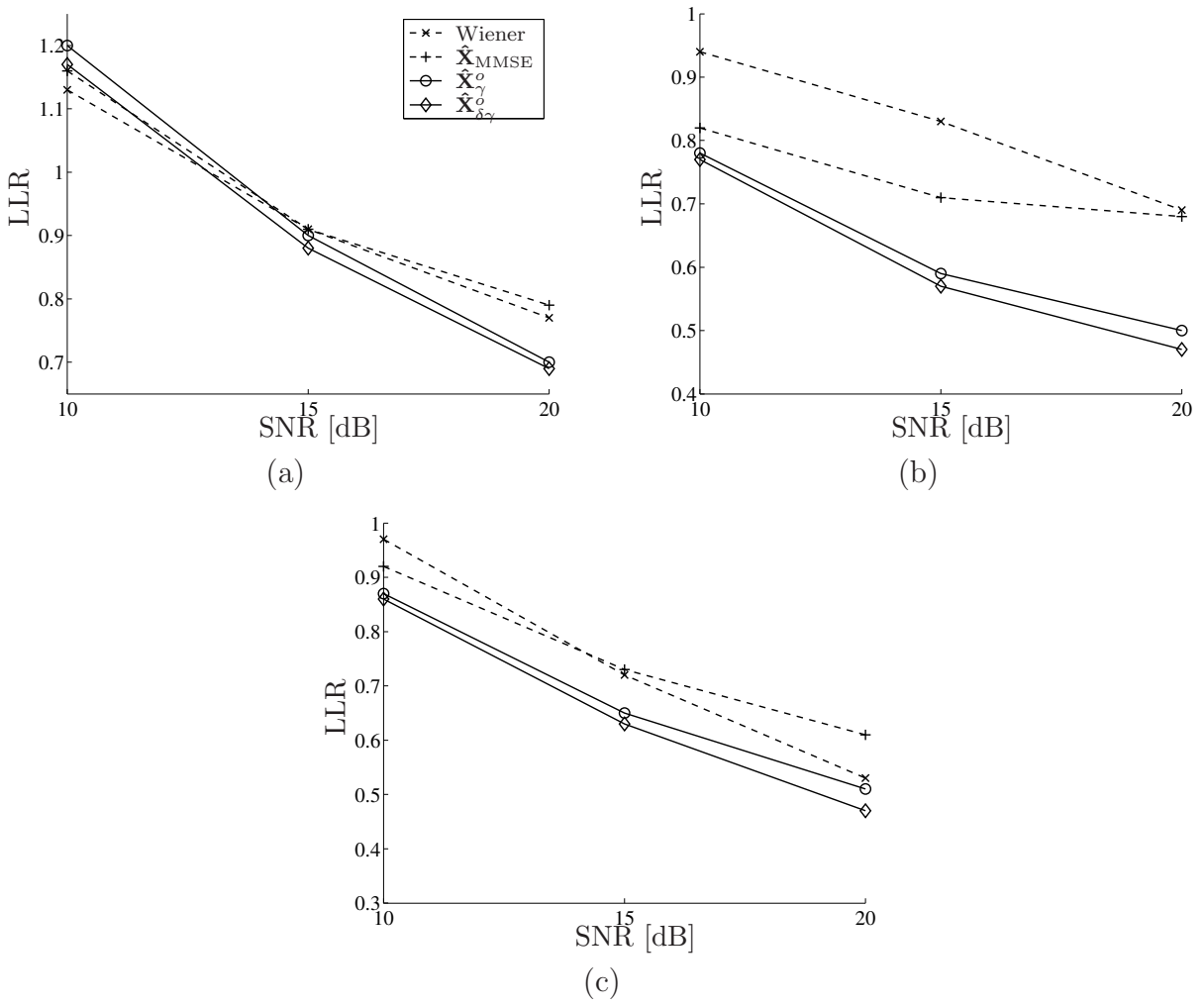

Fig. 8.5 LLR values versus SNR for (a) white noise (b) pink noise (c) aircraft cockpit noise.

Fig. 8.5 presents LLR results for the best estimators identified in Table 8.11, i.e., Wiener, $\hat{\mathbf{X}}_{\mathrm{MMSE}}$, $\hat{\mathbf{X}}^o_\gamma$ and $\hat{\mathbf{X}}^o_{\delta\gamma}$ for white, pink and aircraft cockpit noises. As can be observed, the comparison between the existing and the proposed algorithms were quite different between white and colored noises (i.e. pink and cockpit). In fact, for white noise with an SNR of 20 dB, the proposed estimators gave the best results while for the 10 dB case, the Wiener and $\hat{\mathbf{X}}_{\mathrm{MMSE}}$ were slightly better than the proposed estimators. For the colored

noise cases, the proposed estimators were always better.

8.5.4 Discussion

As mentioned previously, only the phase differs between $\hat{\mathbf{X}}_{\text{MMSE}}$ (7.13) and $\hat{\mathbf{X}}_L^o$ (7.10, 7.16). While $\hat{\mathbf{X}}_{\text{MMSE}}$ has an optimal phase in the sense of the MMSE estimator of $\mathbf{X}$, $\hat{\mathbf{X}}_L^o$ uses the phase of the noisy speech. The phase of the noisy speech was found to be an optimal estimator of the clean speech phase for the uncorrelated frequency component case [4]. We computed for several utterances the differences between the phase of $\hat{\mathbf{X}}_{\text{MMSE}}$ and the clean speech phase and also between the phase of $\hat{\mathbf{X}}_L^o$ (which is the phase of the noisy speech) and the clean speech phase. We found that on average the phase of the noisy speech is closer to the phase of the clean speech than the phase of $\hat{\mathbf{X}}_{\text{MMSE}}$. Therefore, $\hat{\mathbf{X}}_L^o$ should be a better estimator of the clean speech than $\hat{\mathbf{X}}_{\text{MMSE}}$ as observed in Table 8.11. Nevertheless, only small perceptual differences were identified in the informal listening experiments.

We also experimented with shorter window lengths (i.e. $N = 128$ and $N = 256$) for white noise at 20 dB. The results showed that the advantages of the proposed estimators over the existing ones were slightly less for these shorter windows than for $N = 512$. Moreover, the use of the soft threshold structured estimator $\hat{\mathbf{R}}_{\mathbf{X},i}^{\delta}$ did not yield any advantage over the use of the unstructured $\hat{\mathbf{R}}_{\mathbf{X},i}$ for these shorter window lengths. It is not straightforward to identify what causes those different results as the estimation of $\mathbf{R}_{\mathbf{X}}$ is also greatly affected by a shorter window which in return affects the quality of the associated estimator.

Only results for the value of $\gamma = 0.5$ in the $\hat{\mathbf{X}}_{\gamma}^o$ and $\hat{\mathbf{X}}_{\delta\gamma}^o$ estimators were reported. However, we also performed experiments with other values of γ. As expected, the results indicated that choosing values for γ closer to 0 yielded an enhanced speech closer to the one obtained with $\hat{\mathbf{X}}_L^o$ while choosing a value closer to 1 yielded an enhanced speech closer to $\hat{\mathbf{X}}_U^o$.

Wideband PESQ and LLR results were presented here for the 10 dB, 15 dB and 20 dB cases. However, experiments were also conducted for the 0 dB case for which the results are not presented here. For the white and pink noises, the Wiener estimator was found to be superior to the proposed estimators for both LLR and wideband PESQ. However, for the aircraft cockpit noise, the proposed estimator remained better than Wiener for both measures.

In summary, results of informal listening experiments, wideband PESQ and LLR demonstrate noticeable advantages of the proposed estimators over existing ones such as Wiener and MMSE STSA especially at higher SNR values. In particular, the estimator $\hat{\mathbf{X}}^{o}_{\delta\gamma}$ offers a good compromise between speech quality and background noise quantity and whiteness.

Chapter 9

Conclusion

9.1 Summary of the work

There are several systems where the removal of background additive noise in a speech signal is desirable. These include mobile phones [7–9], speech coders [10–12], automatic speech recognition systems [13, 14] and hearing aids [15–18]. Over the years, many speech enhancement approaches have been proposed to remove additive noise including the spectral subtraction [2, 24, 25, 27], Wiener [3, 41, 70], subspace [33–36, 136] and Bayesian STSA [4, 6, 28–31] approaches. The latter was found, in a subjective comparison of these different speech enhancement methods, to perform in general better than the other ones [37] in terms of the overall quality of the enhanced speech, the amount of speech distortion introduced by the processing and the background noise reduction.

In the Bayesian estimation approach for single-channel speech enhancement, an estimate of the clean speech is derived, in the frequency domain, by minimizing the expectation of a cost function that penalizes errors in the clean speech estimate. The well known MMSE STSA estimator is obtained when the chosen cost function is the squared error between the

estimated and actual clean speech STSA [4]. Variants of this estimator were also proposed including the LSA [28], β-SA [29] or WE [6] estimators. This thesis analyzed existing single-channel Bayesian STSA estimators for speech enhancement with the aim of proposing new cost functions and statistical models to improve the performance of such estimators and to gain better understanding of their properties.

The analysis of the β-SA estimator performed in [29] provided insight into its operation but only considered the case where the parameter β is positive. Moreover, it relied on empirical observations in establishing a link between the β-SA and LSA estimators. In Chapter 4, we extended the scope of the analysis in [29] to address the above limitations. We first showed that negative values of β had a normalization effect on the original β-SA cost function, therefore resulting in a behavior similar to the WE estimator with its parameter $p < 0$. Moreover, decreasing β below 0 was found to produce an increase in the noise reduction and speech distortion, therefore enabling an extension of the trade-off between speech distortion and noise reduction, as compared to the strictly positive β case of [29]. Finally, we proved mathematically that the case $\beta \to 0$ indeed corresponds to the LSA estimator. The β-SA estimator with negative β values in the range of $-2 < \beta < 0$ was evaluated experimentally. It was shown that the β-SA estimator with $\beta = -1$ slightly outperforms the well known MMSE STSA and LSA estimators in terms of PESQ while the overall informal MOS appreciation was found to be better than both MMSE STSA and LSA for white noise.

As described in Chapter 3, the WE estimator [6] incorporates a weighting factor while the β-SA estimator [29] incorporates a power law. The parameters accounting for these effects can be given perceptual interpretations that were not considered in [6,29]. In Chapter 5, we first derived and analyzed a new family of Bayesian STSA estimators, referred to as the Wβ-SA estimators, that combined the power law of the β-SA estimator and the weighting

factor of the WE estimator. The parameters (i.e. β and α) entering in the gain function of the Wβ-SA estimator were chosen according to characteristics of the human auditory system, namely, the compressive nonlinearities of the cochlea, the perceived loudness and the ear's masking properties. It was found that choosing the parameters in this way results in a decrease of the estimator gain at high frequencies. This frequency dependence of the gain improved the noise reduction while limiting the speech distortion. The Wβ-SA family of estimators, with the proposed frequency dependent selection of its parameters, was evaluated and compared against existing estimators. In particular, it was shown, using both objective and subjective performance measures, that the new estimators achieved better enhancement performance, especially at low SNR values, when compared to existing Bayesian STSA estimators such as the MMSE STSA, LSA and WE estimator.

In Chapter 6, we noted that the different cost functions presented in Chapters 3 through 5 all had a structure involving a weighted squared difference between a monotonic function of the estimated and actual clean speech STSA. We therefore proposed an analytical generalization of the corresponding estimators which we termed the GWSA family of estimators. The latter incorporates the parameters present in other existing estimators (e.g. α and β) but also features a new parameter denoted as η. These parameters control the shape of the estimator's gain curve as a function of the instantaneous SNR. In contrast to the other parameters, η acts only on the estimated clean speech STSA. It was found that, for appropriate parameter values, η yields an added flexibility in terms of achievable gain curves when compared to existing Bayesian STSA estimators. Also, we showed that all the estimators belonging to the new estimator family tend to a Wiener filter for high instantaneous SNR. This work thus allowed a unification of several existing Bayesian STSA estimators and, moreover, provided a better understanding of this general class of estimators.

In the Bayesian STSA estimators of Chapters 3 through 6, the spectral components are always assumed uncorrelated. However, this assumption is inexact since some correlation is present in practice. In Chapter 7, we thus investigated a multi-dimensional Bayesian STSA estimator that assumed correlated frequency components. Since the closed-form solution of this optimum estimator is not readily available, we alternatively derived closed-form expressions for an upper and a lower bound on the desired estimator. Using these bounds, we proposed a new family of speech enhancement estimators that are characterized by a scalar parameter $0 \leq \gamma \leq 1$, with $\gamma = 0$ corresponding to the lower bound and $\gamma = 1$ to the upper bound. We compared the proposed estimators with the traditional Wiener and MMSE STSA estimators, i.e. that both consider uncorrelated frequency components, as well as with an MMSE estimator of the complex STFT coefficients that assumes correlated frequency components. Results using the wideband extension of the PESQ and LLR measures as well as informal listening experiments showed that the proposed estimators can achieve better performances than benchmarked estimators for several noise types and SNR conditions.

9.2 Future research

Promising avenues for future research have emerged based on the work presented in this Thesis. These are summarized briefly below:

1. *Frame-based adaptive β value*: As mentioned in Chapter 3, it was proposed in [29] to adapt the value of β, in the β-SA estimator, according to each frame's SNR. In Chapter 5, we proposed another approach were β is chosen based on auditory considerations. It would be interesting to investigate how these two approaches can be combined to find optimal values of β. This may indeed result in still better

enhancement performance.

2. *Improved phase estimation in the multi-dimensional STSA estimator allowing for correlated frequency components*: In Chapter 7, the noisy phase was used as an estimator of the clean speech phase and combined with the multidimensional STSA estimator to obtain an estimator of the STFT coefficients. We could alternatively find an optimal multi-dimensional phase estimator that allows for correlated frequency components. It could be that, as in the scalar case [4], the optimal phase estimator is indeed the noisy phase.

3. *Multi-dimensional estimator allowing for correlated STSA but considering uncorrelated phase*: In Chapter 7, we mentioned that the STSA were observed to be more strongly correlated than the STFT. It would therefore be relevant to investigate an estimator that would consider the STSA as correlated but the phase as uncorrelated. This kind of estimator may yield better results than the family of estimators presented in Chapter 7 which considered the complex STFT components as correlated.

4. *Frame-based selection of parameter γ in the multi-dimensional STSA estimator allowing for correlated frequency components*: In the multi-dimensional Bayesian STSA estimator of Chapter 7, the parameter γ allows to adjust the relative weights of the upper and lower bounds in the proposed family of STSA estimator. In our work, a fixed value of $\gamma = 0.5$ was used. One interesting research avenue would be to choose γ adaptively for each frame, e.g. based on SNR considerations, to obtain an estimator closer to the desired one as given by (7.9).

5. *Auditory based generalization of the multi-dimensional STSA estimator*: It would be relevant to derive multi-dimensional Bayesian STSA estimators, such as those

in Chapter 7 that allow for correlated frequency components but, moreover, that make use of more general forms of cost functions, similar to the β-SA or WE cost functions used in the scalar case. It should then be possible to incorporate auditory based features in the multi-dimensional STSA estimator. As in the scalar case, it is expected that this type of approach may result in superior enhancement performance.

9.3 Final remark

The work in this thesis has led to many interesting developments in speech enhancement. The different estimators proposed showed significant performance improvements over existing estimators. However, the level of improvement varied depending on the noise types and SNR levels of the noisy speech. A well engineered solution to the problem of speech enhancement will therefore most likely be formed of many different approaches and not of one single elusive scheme.

Appendix A

Additional derivations for the multi-dimensional estimator

In this appendix, we develop the solutions to the integrals in (7.28) and show that it yields (7.29). We start by solving for the numerator of (7.28), which we denote by N_k. The latter can be written as a product and sum of single integrals

$$N_k = \sum_{\substack{r=0 \\ r\neq t}}^{N-1}\sum_{t=0}^{N-1} U_{kt}^* U_{kr} \prod_{\substack{m=0 \\ m\neq r,t}}^{N-1} \left(\int g(V_m)dV_m\right) \int V_t^* g(V_t)dV_t \int V_r g(V_r)dV_r$$
$$+ \sum_{p=0}^{N-1} |U_{kp}|^2 \prod_{\substack{m=0 \\ m\neq p}}^{N-1} \left(\int g(V_m)dV_m\right) \int |V_p|^2 g(V_p)dV_p \tag{A.1}$$

We need to evaluate four different integrals in (A.1). In order to integrate on real variables instead of complex ones, we will perform the following change of variables : $V_l = v_l e^{j\beta_l}$. The Jacobian associated with that change of variable is $J = v_l$. Let us evaluate the

first integral in (A.1):

$$\begin{aligned}\int g(V_m)dV_m &= \int e^{\{\tilde{Y}_m^* V_m + V_m^* \tilde{Y}_m - |V_m|^2 \lambda_m\}} dV_m \\ &= \int_0^\infty v_m e^{-v_m^2 \lambda_m} \int_{-\pi}^{\pi} e^{2\tilde{y}_m v_m \cos(\beta_m - \measuredangle \tilde{Y}_m)} d\beta_m dv_m \\ &= 2\pi \int_0^\infty v_m e^{-v_m^2 \lambda_m} J_0(-2j\tilde{y}_m v_m) dv_m \\ &= \pi \lambda_m^{-1} M(1, 1; \tilde{y}_m^2/\lambda_m)\end{aligned} \tag{A.2}$$

where $\tilde{Y}_m = \tilde{y}_m e^{j\measuredangle \tilde{Y}_m}$, $J_n(\cdot)$ is a Bessel function of the first kind, $M(a, b; c)$ is the confluent hypergeometric function [83] and (6.631.1) of [83] was used in the last line. The second integral can be evaluated similarly as:

$$\begin{aligned}\int V_t^* g(V_t)dV_t &= \int V_t^* e^{\{\tilde{Y}_t^* V_t + V_t^* \tilde{Y}_t - |V_t|^2 \lambda_t\}} dV_t \\ &= \int_0^\infty v_t^2 e^{-v_t^2 \lambda_t} \int_{-\pi}^{\pi} e^{-j\beta_t + 2\tilde{y}_t v_t \cos(\beta_t - \measuredangle \tilde{Y}_t)} d\beta_t dv_t \\ &= 2\pi j e^{-j\measuredangle \tilde{Y}_t} \int_0^\infty v_t^2 e^{-v_t^2 \lambda_t} J_1(-2j\tilde{y}_t v_t) dv_t \\ &= \pi \tilde{Y}_t^* \lambda_t^{-2} M(2, 2; \tilde{y}_t^2/\lambda_t).\end{aligned} \tag{A.3}$$

The third integral is the complex conjugate of (A.3) and can be similarly shown to be:

$$\int V_r g(V_r)dV_r = \pi \tilde{Y}_r \lambda_r^{-2} M(2, 2; \tilde{y}_r^2/\lambda_r). \tag{A.4}$$

The first integral of the second summation in (A.1) is already given by (A.2) while the

second integral of the second summation can be evaluated as:

$$\begin{aligned}\int |V_p|^2 e^{\{\tilde{Y}_p^* V_p + V_p^* \tilde{Y}_p - |V_p|^2 \lambda_p\}} dV_p &= \int_0^\infty v_p^3 e^{-v_p^2 \lambda_p} \int_{-\pi}^{\pi} e^{2\tilde{y}_p v_p \cos(\beta_p - \measuredangle \tilde{Y}_p)} d\beta_p dv_p \\ &= 2\pi \int_0^\infty v_p^3 e^{-v_p^2 \lambda_p} J_0(-2j\tilde{y}_p v_p) dv_p \\ &= \pi \lambda_p^{-2} M(2, 1; \tilde{y}_p^2/\lambda_p).\end{aligned} \tag{A.5}$$

We can therefore replace (A.2), (A.3), (A.4) and (A.5) in (A.1) to get:

$$\begin{aligned}N_k = \pi^N \sum_{\substack{r=0 \\ r\neq t}}^{N-1} \sum_{t=0}^{N-1} U_{kt}^* U_{kr} \prod_{\substack{m=0 \\ m\neq r,t}}^{N-1} \left(\lambda_m^{-1} M(1, 1; \tilde{y}_m^2/\lambda_m)\right) \frac{\tilde{Y}_t^* \tilde{Y}_r}{\lambda_t^2 \lambda_r^2} M(2, 2; \tilde{y}_t^2/\lambda_t) M(2, 2; \tilde{y}_r^2/\lambda_r) \\ + \pi^N \sum_{p=0}^{N-1} |U_{kp}|^2 \prod_{\substack{m=0 \\ m\neq p}}^{N-1} \left(\lambda_m^{-1} M(1, 1; \tilde{y}_m^2/\lambda_m)\right) \lambda_p^{-2} M(2, 1; \tilde{y}_p^2/\lambda_p)\end{aligned} \tag{A.6}$$

Using (9.212.1) from [83], i.e.:

$$M(\alpha, \gamma; z) = e^z M(\gamma - \alpha, \gamma; -z) \tag{A.7}$$

and the fact that $M(0, \gamma; z) = 1$, we get:

$$N_k = \pi^N \prod_{m=0}^{N-1} \left(\lambda_m^{-1} M(1, 1; \tilde{y}_m^2/\lambda_m)\right) \left(\sum_{\substack{r=0 \\ r\neq t}}^{N-1} \sum_{t=0}^{N-1} U_{kt}^* U_{kr} \frac{\tilde{Y}_t^* \tilde{Y}_r}{\lambda_t \lambda_r} + \sum_{p=0}^{N-1} \frac{|U_{kp}|^2}{\lambda_p} M(-1, 1; -\tilde{y}_p^2/\lambda_p) \right) \tag{A.8}$$

Now using the fact that $M(-1, 1; -\tilde{y}_p^2/\lambda_p) = \tilde{y}_p^2/\lambda_p + 1$ in the last line, we get:

$$N_k = \pi^N \prod_{m=0}^{N-1} \left(\lambda_m^{-1} M(1, 1; \tilde{y}_m^2/\lambda_m)\right) \left(\sum_{r=0}^{N-1} \sum_{t=0}^{N-1} U_{kt}^* U_{kr} \frac{\tilde{Y}_t^* \tilde{Y}_r}{\lambda_t \lambda_r} + \sum_{p=0}^{N-1} \frac{|U_{kp}|^2}{\lambda_p} \right) \tag{A.9}$$

Next let us evaluate the denominator in (7.28). We notice that the integral in the latter is identical to (A.2), therefore:

$$\prod_{m=0}^{N-1} \int g(V_m) dV_m = \pi^N \prod_{m=0}^{N-1} \lambda_m^{-1} M(1, 1; \tilde{y}_m^2/\lambda_m). \tag{A.10}$$

Combining (A.9) and (A.10) in (7.28), we get the following:

$$E\{\mathcal{X}_k^2|\mathbf{Y}\} = \sum_{r=0}^{N-1} \sum_{t=0}^{N-1} U_{kt}^* U_{kr} \frac{\tilde{Y}_t^* \tilde{Y}_r}{\lambda_t \lambda_r} + \sum_{p=0}^{N-1} \frac{|U_{kp}|^2}{\lambda_p}$$

which is (7.29).

Appendix B

Harvard sentences used in experiments

The following Harvard sentences were used in the different experiments included in this thesis [122]. The list and sentence numbers refer to the ones of [122]. Each list was spoken by a different individual, therefore, 3 males and 3 females spoke each 5 sentences.

Male spoken utterances:

- List 7
    1. We talked of the side show in the circus.
    2. Use a pencil to write the first draft.
    3. He ran half way to the hardware store.
    4. The clock struck to mark the third period.
    5. A small creek cut across the field.

- List 13
    1. Type out three lists of orders.
    2. The harder he tried the less he got done.
    3. The boss ran the show with a watchful eye.
    4. The cup cracked and spilled its contents.
    5. Paste can cleanse the most dirty brass.

- List 19
    1. Acid burns holes in wool cloth.
    2. Fairy tales should be fun to write.
    3. Eight miles of woodland burned to waste.
    4. The third act was dull and tired the players.
    5. A young child should not suffer fright.

Female spoken utterances:

- List 1
    1. The birch canoe slid on the smooth planks.
    2. Glue the sheet to the dark blue background.
    3. It's easy to tell the depth of a well.
    4. These days a chicken leg is a rare dish.
    5. Rice is often served in round bowls.

- List 25
    1. On the islands the sea breeze is soft and mild.
    2. The play began as soon as we sat down.
    3. This will lead the world to more sound and fury.
    4. Add salt before you fry the egg.
    5. The rush for funds reached its peak Tuesday.

- List 31
    1. Slide the box into that empty space.
    2. The plant grew large and green in the window.
    3. The beam dropped down on the workmen's head.
    4. Pink clouds floated with the breeze.
    5. She danced like a swan, tall and graceful.

References

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Washington, DC), pp. 208 – 211, 1979.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 113–120, Apr. 1979.

[3] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, pp. 1109–1121, Dec. 1984.

[5] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech, Audio Process.*, vol. 5, pp. 497–514, Nov. 1997.

[6] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech, Audio Process.*, vol. 13, pp. 857–869, Sep. 2005.

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

[8] M. M. Goulding and J. S. Bird, "Speech enhancement for mobile telephony," *IEEE Trans. Vehicul. Tech.*, vol. 39, pp. 316–326, Nov. 1990.

[9] C. H. You, S. N. Koh, and S. Rahardja, "Adaptive β-order MMSE speech enhancement application for mobile communication in a car environment," in *Proc. Joint Conf. 4th Int. Conf. on Information, Communications and Signal Processing and 4th Pacific Rim Conf. on Multimedia*, pp. 1629–1632, 2003.

[10] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech, Audio Process.*, vol. 3, pp. 59–71, Jan. 1995.

[11] R. Martin and R. V. Cox, "New speech enhancement techniques for low bit rate speech coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 165–167, Jun. 1999.

[12] T. Agarwal and P. Kabal, "Pre-processing of noisy speech for voice coders," in *Proc. IEEE Speech Coding Workshop*, (Tsukuba City, Japan), pp. 169–171, 2002.

[13] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Process.*, vol. 39, pp. 795–805, Apr 1991.

[14] T. Yamada, M. Kumakura, and N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 2006–2013, Nov. 2006.

[15] L. Min, H. G. McAllister, N. D. Black, and T. A. De Perez, "Perceptual time-frequency subtraction algorithm for noise reduction in hearing aids," *IEEE Trans. Biomed. Eng.*, vol. 48, pp. 979–988, Sep. 2001.

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

[17] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wooters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Trans. Signal Process.*, vol. 55, pp. 1579–1585, Apr. 2007.

[18] O. Roy and M. Vetterli, "Rate-constrained collaborative noise reduction for wireless hearing aids," *IEEE Trans. Signal Process.*, vol. 57, pp. 645–657, Feb. 2009.

[19] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 177–180, 1987.

[20] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech, Audio Process.*, vol. 6, pp. 373–385, Jul. 1998.

[21] 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, Language Process.*, vol. 14, pp. 19–32, Jan. 2006.

[22] M. R. Schroeder, "U.S. patent 3180936: Apparatus for supressing noise and distortion in communication signals," Apr. 1965.

[23] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, pp. 471–472, Oct. 1978.

[24] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech, Audio Process.*, vol. 7, pp. 126–137, Mar. 1999.

[25] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech, Audio Process.*, vol. 9, pp. 799–807, Nov. 2001.

[26] S. D. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Orlando, FL), 2002.

[27] K. Hasan, S. Salahuddin, and R. Khan, "A modified *a priori* SNR for speech enhancement using spectral subtraction rules," *IEEE Signal Process. Lett.*, vol. 11, pp. 450–453, Apr. 2004.

[28] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, pp. 443–445, Apr. 1985.

[29] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech, Audio Process.*, vol. 13, pp. 475–486, Jul. 2005.

[30] J. H. L. Hansen, V. Radhakrishnan, and K. Hoberg Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 2049–2063, Nov. 2006.

[31] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, pp. 1741–1752, Aug. 2007.

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

[33] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Process.*, vol. 3, pp. 251–266, Jul. 1995.

[34] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Process.*, vol. 11, pp. 700–708, Nov. 2003.

[35] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech, Audio Process.*, vol. 11, pp. 334–341, Jul. 2003.

[36] C. H. You, S. N. Koh, and S. Rahardja, "An invertible frequency eigendomain transformation for masking-based subspace speech enhancement," *IEEE Signal Process. Lett.*, vol. 12, pp. 461–464, May 2005.

[37] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, pp. 588–601, 2007.

[38] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes, 4th Edition*. McGraw-Hill, 2002.

[39] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech, Audio Process.*, vol. 11, pp. 845–856, Sep. 2005.

[40] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, pp. 679–681, Aug. 1982.

[41] R. J. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, pp. 137–145, Apr. 1980.

[42] P. E. Papamichalis, *Practical Approaches to Speech Coding*. Prentice Hall, 1987.

[43] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*. Prentice Hall, 1988.

[44] ITU-T, *Recommendation P.862: Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*. Feb. 2001.

[45] ITU-T, *Recommendation P.800: Methods for Subjective Determination of Transmission Quality*. Aug. 1996.

[46] ITU-R, *Recommendation BS.1534-1: Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*. 2001.

[47] S. Saito and K. Nakata, *Fundamentals of Speech Signal Processing*. Academic Press, 1985.

[48] J. Benesty, M. Sondhi, and Y. Huang, eds., *Handbook of Speech Processing*. Springer, 2008.

[49] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, 2002.

[50] D. O'Shaughnessy, *Speech Communications: Human and Machine, 2nd Edition*. IEEE Press, 2000.

[51] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, *Principles of Neural Science, 4th Edition*. McGraw-Hill, 2000.

[52] P. Brodal, *The Central Nervous System*, ch. "The auditory system", pp. 208–220. Oxford University Press, Third ed., 2004.

[53] L. Robles and M. A. Ruggero, "Mechanics of the mammalian cochlea," *Physiological Rev.*, vol. 81, pp. 1305–1352, Jul. 2001.

[54] B. C. J. Moore, *An Introduction to the Psychology of Hearing, 5th Edition*. Academic Press, 2004.

[55] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models, 2nd Edition*. Information Sciences, Springer, 1999.

[56] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inf. Theory*, vol. 38, pp. 824–839, Mar. 1992.

[57] B. Moore, D. Vickers, C. Plack, and A. Oxenham, "Inter-relationship between different psychoacoustic measures assumed to be related to the cochlear active mechanism," *J. Acoust. Soc. Am.*, vol. 106, pp. 2761–2778, 1999.

[58] E. A. Lopez-Poveda, C. J. Plack, and R. Meddis, "Cochlear nonlinearity between 500 and 8000 Hz in listeners with normal hearing," *J. Acoust. Soc. Am.*, vol. 113, pp. 951–960, Feb. 2003.

[59] C. J. Plack and C. G. O'Hanlon, "Forward masking additivity and auditory compression at low and high frequencies," *Journal of the Association for Research in Otholaryngology*, vol. 4, pp. 405–415, Sep. 2003.

[60] P. S. Rosengard, A. J. Oxenham, and L. D. Braida, "Comparing different estimates of cochlear compression in listeners with normal and impaired hearing," *J. Acoust. Soc. Am.*, vol. 117, pp. 3028–3041, May 2005.

[61] Y. Ephraim, H. Lev-Ari, and W. J. J. Roberts, *The Electronics Handbook, 2nd Edition*, ch. "A brief survey of speech enhancement". CRC Press, 2005.

[62] 3GPP2, *Enhanced Variable Rate Codec, Speech Service Options 3, 68, and 70 for Wideband Spread Spectrum Digital Systems*. Sep. 2006.

[63] ITU-T, *Recommendation G.711.1: Wideband Embedded Extension for G.711 Pulse Code Modulation*. Mar. 2008.

[64] R. Le Bouquin Jeannes, P. Scalart, G. Faucon, and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: a survey," *IEEE Trans. Speech, Audio Process.*, vol. 9, pp. 808–820, Nov. 2001.

[65] Y. Hu and P. C. Loizou, "A comparative intelligibility study of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Honolulu, HI), pp. 561–564, 2007.

[66] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech, Audio Process.*, vol. 10, pp. 341–351, Sep. 2002.

[67] P. Kabal, *Windows for Transform Processing*. Tech. Rep., McGill University, 2005. http://www-mmsp.ece.mcgill.ca/Documents/Reports/2005/KabalR2005a.pdf.

[68] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, pp. 236–243, Apr. 1984.

[69] J. B. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, pp. 235–238, Jun. 1977.

[70] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 1218–1234, Jul. 2006.

[71] H. Ding, I. Y. Soon, S. N. Koh, and C. K. Yeo, "A spectral filtering method based on hybrid wiener filters for speech enhancement," *Speech Commun.*, vol. 51, pp. 259–267, 2009.

[72] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, pp. 113–116, Apr. 2002.

[73] I. Cohen, "Speech enhancement using a noncausal *a priori* SNR estimator," *IEEE Signal Process. Lett.*, vol. 11, pp. 725–728, Sep. 2004.

[74] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. Speech, Audio Process.*, vol. 13, pp. 870–881, Sep. 2005.

[75] R. C. Hendriks, R. Heusdens, and J. Jensen, "An MMSE estimator for speech enhancement under a combined stochastic-deterministic speech model," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, pp. 406–415, Feb. 2007.

[76] J. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech Commun.*, vol. 49, pp. 530–541, 2007.

[77] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[78] H. L. Van Trees, *Detection, Estimation and Modulation Theory: Part I*. Wiley, 1968.

[79] I. Andrianakis and P. R. White, "Speech spectral amplitude estimators using optimally shaped gamma and chi priors," *Speech Commun.*, vol. 51, pp. 1–14, Jan. 2009.

[80] S. Gazor and W. Zhang, "Speech enhancement employing Laplacian-Gaussian mixture," *IEEE Trans. Speech, Audio Process.*, vol. 13, pp. 896–904, Sept. 2005.

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

[82] B. J. Shannon and K. K. Paliwal, "Role of phase estimation in speech enhancement," in *Proc. 9th Int. Conf. Spoken Language Processing - Interspeech*, (Pittsburgh, PA), pp. 1423–1426, 2006.

[83] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products, 6th Edition*. Academic Press, 2000.

[84] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP J. Appl. Signal Process.*, vol. 10, pp. 1043–1051, 2003.

[85] Y. Ephraim and I. Cohen, *The Electrical Engineering Handbook, 3rd Edition*, ch. "Recent advancements in speech enhancement". CRC Press, 2005.

[86] C. H. You, S. N. Koh, and S. Rahardja, "An MMSE speech enhancement approach incorporating masking properties," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Montreal, Canada), pp. 725–728, 2004.

[87] A. H. Gray Jr. and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, pp. 380–391, Oct. 1976.

[88] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech, Audio Process.*, vol. 2, pp. 345–349, Apr. 1994.

[89] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 2098–2108, Nov. 2006.

[90] J. Sohn and N. S. Kim, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, pp. 1–3, Jan. 1999.

[91] Y. D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Process. Lett.*, vol. 8, pp. 276–278, Oct. 2001.

[92] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, pp. 271–287, Apr. 2004.

[93] J. M. Gorriz, J. Ramirez, E. W. Lang, and C. G. Puntonet, "Jointly Gaussian PDF-based likelihood ratio test for voice activity detection," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, pp. 1565–1578, Nov. 2008.

[94] V. Stahl, A. Fischer, and R. Bippus, "Quatile based noise estimation for spectral subtraction and wiener filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1875–1878, 2000.

[95] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech, Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[96] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 2, pp. 12–15, 2002.

[97] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[98] J. S. Erkelens and R. Heusdens, "Tracking of nonstationary noise based on data-driven recursive noise power estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, pp. 1112–1123, Aug. 2008.

[99] E. Plourde and B. Champagne, "Further analysis of the β-order MMSE STSA estimator for speech enhancement," in *Proc. 20th IEEE Canadian Conf. on Electrical and Computer Eng.*, (Vancouver, Canada), pp. 1594–1597, 2007.

[100] E. Plourde and B. Champagne, "Auditory based spectral amplitude estimators for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, pp. 1614–1623, Nov. 2008.

[101] P. J. Wolfe and S. J. Godsill, "Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Istanbul, Turkey), pp. 821–824, 2000.

[102] P. J. Wolfe and S. J. Godsill, "A perceptually balanced loss function for short-time spectral amplitude estimation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Hong Kong), pp. 425–428, 2003.

[103] R. Meddis and L. P. O'Mard, "A computational algorithm for computing nonlinear auditory frequency selectivity," *J. Acoust. Soc. Am.*, vol. 109, pp. 2852–2861, Jun. 2001.

[104] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Washington : U.S. Govt. Print. Off., 1964.

[105] T. L. Petersen and S. F. Boll, "Acoustic noise suppression in the context of a perceptual model," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Atlanta, GA), pp. 1086–1088, 1981.

[106] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, Apr. 1990.

[107] D. D. Greenwood, "A cochlear frequency-position function for several species - 29 years later," *J. Acoust. Soc. Am.*, vol. 87, pp. 2592–2605, Jun. 1990.

[108] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *J. Audio Eng. Soc.*, vol. 45, pp. 224–240, Apr. 1997.

[109] C. Formby and R. B. Monsen, "Long-term average speech spectra for normal and hearing-impaired adolescents," *J. Acoust. Soc. Am.*, vol. 71, no. 1, pp. 196–202, 1982.

[110] E. Plourde and B. Champagne, "Integrating the cochlea's compressive nonlinearity in the bayesian approach for speech enhancement," in *Proc. 15th European Signal Processing Conf.*, (Poznań, Poland), pp. 70–74, 2007.

[111] E. Plourde and B. Champagne, "Perceptually based speech enhancement using the weighted β-SA estimator," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, (Las Vegas, NV), pp. 4193–4196, 2008.

[112] E. Plourde and B. Champagne, "Generalized Bayesian estimators of the spectral amplitude for speech enhancement," *IEEE Signal Process. Lett.*, vol. 16, pp. 485–488, Jun. 2009.

[113] C. Li and S. V. Andersen, "A block-based linear MMSE noise reduction with a high temporal resolution modeling of the speech excitation," *EURASIP J. Appl. Signal Process.*, vol. 18, pp. 2965–2978, 2005.

[114] U. of Texas at Dallas, "North texas vowel database." [Online] http://www.utdallas.edu/∼assmann/KIDVOW1/North_Texas_vowel_database.html.

[115] D. Sarason, *Complex Function Theory, 2nd Edition*. American Mathematical Society, 2007.

[116] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory*. Princeton University Press, 2005.

[117] W. Rudin, *Real and Complex Analysis, 3rd Edition*. McGraw-Hill, 1987.

[118] S. L. Marple, *Digital Spectral Analysis*. Prentice Hall, 1987.

[119] E. Plourde and B. Champagne, "Multi-dimensional Bayesian STSA estimators for the enhancement of speech with correlated frequency components," *IEEE Trans. Audio, Speech, Language Process.*, pp. 1–9, submitted, Aug. 2009.

[120] E. Plourde and B. Champagne, "Bayesian spectral amplitude estimation for speech enhancement with correlated spectral components," in *Proc. 2009 IEEE Workshop on Statistical Signal Processing*, (Cardiff, U.K.), 2009. Accepted for presentation.

[121] Rice University, "Signal processing information base: Noise data." [Online] http://spib.rice.edu/spib/select_noise.html.

[122] "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, Sep. 1969.

[123] ITU-T, *Recommendation P.56: Objective Measurement of Active Speech Level*. Mar. 1993.

[124] P. Kabal, *Measuring Speech Activity*. Tech. Rep., McGill University, 2000.

[125] V. Grancharov, *Human Perception in Speech Processing*. PhD thesis, KTH (Royal Institute of Technology), 2006.

[126] ITU-T, *Recommendation P.835: Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithm*. Nov. 2003.

[127] V. Grancharov, J. Samuelsson, and B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 764–773, May 2006.

[128] E. Vincent, "MUSHRAM: A MATLAB interface for MUSHRA listening tests." [Online] http://www.elec.qmul.ac.uk/people/emmanuelv/mushram/.

[129] R. Huber and B. Kollmeier, "PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 1902–1911, Nov. 2006.

[130] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. 5th Int. Conf. Spoken Language Processing*, (Sydney, Australia), pp. 2819–2822, 1998.

[131] ITU-T, *Recommendation P.862.2: Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*. Nov. 2005.

[132] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. 9th Int. Conf. Spoken Language Processing - Interspeech*, (Pittsburgh, PA), pp. 1447–1450, 2006.

[133] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, pp. 229–238, Jan. 2008.

[134] S. Zhang and J. Jin, *Computation of special functions*. John Wiley, 1996.

[135] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers, 3rd Edition*. John Wiley & Sons, 2003.

[136] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. Speech, Audio Process.*, vol. 8, pp. 497–507, Sep. 2000.