Perceptual and Multi-Microphone Signal Subspace Techniques for Speech Enhancement

Firas Jabloun

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

October 2004

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

To my parents, sisters, brother and to my wife

Abstract

The performance of speech communication systems, such as hands-free telephony, is known to seriously degrade under adverse acoustic environments. The presence of noise can lead to the loss of intelligibility as well as to the listener's fatigue. These problems can generally make the existing systems unsatisfactory to the customers especially that the offered services usually put no restrictions on where they can actually be used. For this reason, speech enhancement is vital for the overall success of these systems on the market.

In this thesis we present new speech enhancement techniques based on the signal subspace approach. In this approach the input speech vectors are projected onto the signal subspace where it is processed to suppress any remaining noise then reconstructed again in the time domain. The projection is obtained via the eigenvalue decomposition of the speech signal covariance matrix.

The main problem with the signal subspace based methods is the expensive eigenvalue decomposition. In this thesis we present a simple solution to this problem in which the signal subspace filter is updated at a reduced rate resulting in a significant reduction in the computational load. This technique exploits the stationarity of the input speech signal within a frame of 20-30 msec to use the same eigenvalue decomposition for several input vectors. The original implementation scheme was to update the signal subspace filter for every such input vector. The proposed technique was experimentally found to offer significant computational savings at almost no performance side-effects.

The second contribution of this thesis is the incorporation of the human hearing properties in the signal subspace approach using a sophisticated masking model. It is known that there is a tradeoff between the amount of noise reduction achieved and the resulting signal distortion. Therefore, it would be beneficial to avoid suppressing any noise components as long as they are not perceived by the human ear. However, since the masking models available are usually developed in the frequency domain it is not straightforward to use them in the signal subspace framework. Our task consisted in finding a way to map the masking information calculated in the frequency domain into the eigendomain allowing to develop a new perceptual signal subspace method. Subjective tests have supported the claim that the use of the masking criteria has indeed provided the sought performance improvements.

The generalization of the signal subspace approach into a multi-microphone design was also a subject addressed in this thesis. The method we developed takes the form of a con-

ventional beamformer followed by a signal subspace postfilter. The postfilter coefficients are calculated based on the signals gathered from the different available acquisition channels. In the thesis we describe a novel technique to calculate the filter coefficients via averaging in the eigendomain. Simulations show that the new method is insensitive to reverberation time and that it outperforms other competing methods particularly under diffuse noise fields.

To evaluate the performance of the multi-microphone methods, a tool to digitally simulate the room reverberation is needed. To this end, we have presented a generalization of the popular image method into a subband design allowing it to simulate more realistic enclosures where the reflection coefficients of the surfaces are usually frequency dependent. The new subband room simulator readily offers important computational savings thanks to the adopted filterbank design.

Sommaire

Il est connu que la performance des systèmes de communication par la voix se détériore lorsqu'ils sont utilisés dans des environnements acoustiques peu favorables. En effet, la présence du bruit cause la perte de l'intelligibilité et engendre la fatigue chez les auditeurs. Ces problèmes peuvent rendre les systèmes existant sur le marché inintressants pour les clients surtout que les services offerts par les compagnies de télécommunication ne comportent aucune restriction sur les endroits où ils seront utilisés. Dans ce contexte, les algorithmes qui visent à améliorer la qualité du signal parole sont très importants du fait qu'ils permettent à ces systèmes de satisfaire les attentes du marché.

Dans cette thèse, nous présentons des nouvelles techniques, visant à rehausser la qualité de la voix, qui sont basées sur l'approche de sous-espace du signal (SES). Selon cette approche, les vecteurs du signal sont projetés sur le sous-espace du signal où ils sont traités afin d'éliminer le bruit restant. Après ce traitement, les vecteurs seront reconstruits dans le domaine du temps. La projection est obtenue grâce à la décomposition en valeurs propres de la matrice de covariance du signal parole.

Le problème avec l'approche SES est que le coût, en terme de temps de calcul, relié à la décomposition en valeurs propres est élevé. Dans cette thèse, nous proposons une technique simple pour résoudre ce problème. Cette technique réduit considérablement le temps de calcul car le filtre en sous-espace est mis à jour moins fréquemment. Initialement, l'implémentation de l'approche SES consistait à recalculer un nouveau filtre pour chaque vecteur. L'originalité de notre technique réside dans l'exploitation de la stationnarité du signal parole dans un intervalle de 20-30 msec afin d'utiliser la même décomposition en valeurs propres pour plusieurs vecteurs. Les expériences menées montrent que notre nouvelle technique réduit considérablement le coût de calcul tout en conservant la même performance.

La deuxième contribution de la présente thèse est l'incorporation des propriétés auditives humaines dans l'approche de sous-espace du signal en utilisant un modèle de masquage auditif sophistiqué. Il est établi qu'il est difficile de réduire le bruit sans introduire de distorsion. Il serait donc judicieux de l'éliminer seulement quand il est perçu par l'oreille. Cependant puisque tous les modèles de masquage existants sont développés dans le domaine de fréquence, il n'est pas facile d'incorporer les effets de perception dans l'approche de sous-espace du signal. Notre tâche était justement de trouver une façon de transférer l'information du masquage, calculée dans le domaine de fréquence, au domaine des valeurs

propres ce qui permet de concevoir un filtre sous-espace auditif. Les testes subjectifs effectués ont montré que les sujets préfèrent l'application des critères auditif aux méthodes traditionnelles.

La généralisation de la méthode sous-espace du signal à un contexte de multi microphones est aussi un sujet traité dans cette thèse. La méthode développée prend la forme d'un module de formation de voie conventionnelle suivi d'un post-filtre dans le sous-espace signal. Le calcul des coefficients de ce post-filtre est basé sur les signaux obtenus à la sortie des différents microphones disponibles. Dans cette thèse, nous décrivons une nouvelle technique où ces coefficients sont calculés à partir d'une moyenne prise dans le domaine des valeurs propres. Les simulations effectuées ont montré que la nouvelle méthode est insensible au niveau de la réverbération et montre une meilleure performance que les méthodes existantes, surtout quand le bruit peut être modélisé par un champ diffus.

Afin d'évaluer la performance des méthodes multi microphones, il est nécessaire d'avoir un outil de simulation numérique de la réverbération d'une chambre. A cette fin, nous présentons une généralisation en sous-bandes de la populaire méthode des images. La méthode que nous proposons permet de simuler des chambres plus réalistes où les coefficients de réflexion des surfaces peuvent dépendre de la fréquence. Grâce à sa conception le nouveau simulateur de chambre permet en plus de diminuer le temps de calcul.

Acknowledgments

I would like to express my sincere gratitude and appreciation to my supervisor Prof. Benoît Champagne for his support, encouragement and guidance during the course of my Ph.D studies. Without his help and advice, this research would not have been completed.

Special thanks are also extended to Prof. P. Christoffersen, Prof. D. Lowther, Prof. M. Blostein, Prof. R. Rose and Prof. R. Lefebvre for accepting to be on my committee and for providing valuable advice and comments.

The financial support provided by "La Mission Universitaire de la Tunisie" is gratefully acknowledged.

Special thanks to all my friends and to my colleagues in the TSP lab for their warm friendship and for being there for me whenever I needed their help. Particularly I would like to thank Chris and Bryan my friends from Texas for their last minute help towards the initial submission of this thesis and to Souheyl for making my Ph.D experience more enjoyable.

I also want to express my deepest gratitude to my parents, my sisters Jihen and Ines, and to my brother Hassine for their love, support and prayers. Thank you for believing in me.

Very special thanks to my dear wife Ramla for her love, support and patience especially in the last couple of years. From now on we can enjoy normal life :-)

Contents

List of Figures

List of Tables

List of Acronyms

ANC	Adaptive Noise Canceling.
BT	Blackman-Tukey PSD estimator.
CCM	Composite Covariance Matrix.
DCT	Discrete Cosine Transform.
DFT	Discrete Fourier Transform.
DOA	Direction Of Arrival.
DSP	Digital Signal Processing.
DTFT	Discrete Time Fourier Transform.
EVD	EigenValue Decomposition.
FBEVD	Frame-Based EVD implementation.
FET	Frequency to Eigendomain Transformation.
FFT	Fast Fourier Transform.
FLOPS	Floating point Operations.
KLT	Karhunen-Loeve Transform.
LMMSE	Linear Minimum Mean Squared Error.
MEDA	Multi-microphone method with Eigen-Domain Averaging.
MNR	Mask to Noise Ratio.
MSVD	Multi-microphone method with singular value decomposition.
NRF	Noise reduction Factor.
PSD	Power Spectral Density.
PSS	Perceptual Signal Subspace method.
PWSS	Pre-Whitening based Signal Subspace method.
RQSS	Raleigh Quotient Signal Subspace method.
SNR	Signal to Noise Ratio.
SSA	Signal Subspace Approach.
SRS	Subband Room Simulator.
SVD	Singular Value Decomposition.
VAD	Voice Activity Detection.
WSS	Wide Sense Stationary.

Chapter 1

Introduction

For centuries, humans have developed well established sounds to convey a certain desired meaning to a listener hence creating what we call "speech". However, noisy environments hinder the listener's ability to *decode* these sounds and match them with their corresponding meaning, hence degrading the *intelligibility*, i.e. the listener's ability to understand what is being said. Moreover, the presence of sustained noise can cause a significant fatigue or discomfort to the listener due to a reduced *quality*. The latter issue is usually more annoying than the former because most people seem to be unable to tolerate a persistent noise for long time periods. A certain loss in intelligibility, on the other hand, can be tolerated to some extent and the listener usually takes advantage of the context to understand the uttered words and to fill in the missing parts.

Nowadays, in addition to interpersonal communications, speech is also being transmitted via telecommunication channels where the receiver can be either a human or a machine. In such applications, the need to preserve a good speech quality and intelligibility is a vital feature. To this end, intensive research in the area of speech enhancement has been conducted during the last three decades and a variety of methods have been developed. These methods have been found to be very useful in different speech applications.

Yet, due to the complexity of the speech signal, this area of research still poses a considerable challenge. Indeed, any noise reduction technique has to cope with the tradeoff between the amount of noise suppressed and the introduced signal distortion. In this thesis we further investigate this issue and seek to provide novel methods for speech enhancement based on the so called signal subspace approach (SSA) [41].

1.1 Applications of speech enhancement

Recent technological advances in Digital Signal Processing (DSP) have opened new horizons for speech communication applications. Nowadays the services offered to the customers are very broad and at a very high level of sophistication. Nonetheless, the user's expectations are still pushing for even further improvements. Namely, the user is expecting the products offered to be as satisfactory under noisy conditions as they are in quiet. Such demands can only be met by the incorporation, in such systems, of increasingly sophisticated and robust speech enhancement techniques capable to offer the desired performance and to satisfy the user under the most harsh conditions.

In this section we describe some of the most common applications of speech enhancement although the content of this thesis is rather general and is not confined to any particular one of them.

Cellular telephony

The digital mobile technology took the telephone out of its traditional environments to new places like cars, crowded streets, stock markets and even night clubs. These environments are characterized by low Signal-to-Noise Ratio (SNR) which would significantly degrade both the quality and the intelligibility of the transmitted signal. In such situations, a noise reduction module placed between the Analog-to-Digital converter and the encoder would reduce the level of the interfering noise and hence improve the perceived speech signal [30, 52].

For example, for security reasons, some countries have imposed regulations that force the hands-free utilization of cellular telephones in cars. This scenario makes the level of the noise caused by the engine, the wind and the wheels, high enough to significantly disturb any ongoing conversation. For this reason many researchers have tackled this particular problem resulting in several methods that handle car noise [26, 54, 120].

In addition to that, many speech coders rely on a speech model to efficiently compress the signal. These models usually become inaccurate in the presence of noise resulting in an unpleasant signal distortion. A noise reduction module would make this artifact of the coding scheme less annoying.

Teleconferencing

Another area which is gaining popularity is teleconferencing. One reason why teleconferencing is important is that it helps reduce the traveling costs by providing a convenient alternative to group meetings [37] (a feature welcomed by many companies). Unfortunately, teleconferencing systems are typically used in a hands free mode hence making any interference, such as computer fan or air conditioning noises, more annoying.

In this particular application, microphone arrays have been found useful. They are usually designed to pick up a speech signal from one desired look direction and reject interference from other directions [21, 96]. Microphone arrays have also other advantages and capabilities such as dereverberation [1, 111] and automatic localization and tracking of one particular speaker on the fly [13, 14, 122].

Speech recognition

Speech enhancement is also found useful in speech recognition whose applications are constantly gaining popularity. These include for example voice activated command functionality in cars such as hands-free dialing[1]. Retrieving driving direction by means of car navigation systems is another application which becomes more efficient and easier to use via a voice interface. Manufacturers are also embedding speech recognition engines into small devices such as PDA's allowing to access the address book, schedule appointments and send email in the most convenient way. Indeed, software licenses from embedded speech is estimated to increase from US $8 million in 2003 to $227 million in 2006 [102].

However, this relatively new technology still has to face many challenges, the most critical of which seems to be the noise. The presence of noise results in a vast discrepancy between the training and testing conditions. In fact speech recognition engines rely on mapping the speech waveform to a set of features used to train the so-called Hidden Markov Models (HMMs). These features, such as the Mel-Frequency Cepstral Coefficients (MFCC) or the Perceptual Linear Prediction coefficients (PLP), are usually designed in a way that imitates the human auditory system. During recognition, a similar set of features is computed and the best match from the stored patterns is picked and recognized as the uttered word. This technique, although being satisfactory under quiet conditions, drastically fails under noise. For this reason, a noise compensation module is usually included

[1] Important for security reasons.

in the frontend in order to obtain a better match between training and testing conditions of automatic speech recognition (ASR) engines [31]. The use of such modules is found to significantly improve the accuracy of the recognizer [59, 69, 71]. Microphone array based noise reduction methods have also been found useful in ASR applications [4, 98].

Hearing aids

Speech enhancement is also beneficial in applications with a humanitarian motive, namely hearing aids [93]. Users of hearing aids mainly complain about the fact that their equipment amplifies, in addition to the desired speech signal, all the surrounding noises. Many find this drawback so serious that they choose the complete silence to the annoying uncomfortable noisy environment their hearing aid is offering.

Therefore, allowing the hearing aid users to selectively listen to a single sound source, separating it from the surrounding noises, has obvious advantages in this case.

1.2 Speech enhancement methods

Aware of the importance of speech enhancement, researchers have been continually investigating new methods which vary according to their performance, complexity, targeted application and also according to their underlying theoretical principle. We provide below a brief survey of the most popular methods of speech enhancement; a detailed description of those most relevant for this thesis will be presented in Chapter 3.

One possible approach is based on parameter modeling of the speech signal for example with an Auto-Regressive (AR) model [31]. The estimated model parameters are then used to re-synthesize the speech signal [108, 39, 59]. Even though this scheme may result in a reduced noise level, the natural sound of speech is significantly degraded during the synthesis step.

Another more popular class of methods are commonly referred to as the short-term spectral domain methods or the transform domain methods. In general, in these methods, the noisy signal is transformed to an appropriate domain where it is filtered by a usually data dependent filter. This filter suppresses noise by subtracting an estimated noise bias gathered during non-speech activity periods. Finally, an inverse transform is applied to recover the enhanced signal in the time domain. Within this class of methods fall the popular spectral subtraction method [9] and Wiener filtering [31], which are both frequency

domain methods. Another related technique is the signal subspace method which operates in the eigendomain [41, 32].

Transform domain methods usually suffer from an annoying artifact known as the *musical noise*. It is an artificial residual noise resulting from poor estimation of the speech and/or noise parameters used for the design of the suppression filter. Several modifications have been applied to the basic transform domain methods in order to overcome this drawback. One promising approach is the use of the human auditory masking to alter the suppression filter coefficients [157, 150, 56]. The principle of this method is that there is no need to aggressively suppress any noise components as long as they eventually will not be perceived by the human ear. Doing so, a more satisfactory trade off between speech distortion and residual noise level can be achieved.

The use of masking properties has also been suggested in noise reduction methods intended for automatic speech recognition applications. One would argue, however, that this should not have any significant impact since the receiver in this case is a machine rather than a human. Nonetheless, the reported reduction in recognition error rates can be explained as follows. The most popular set of features used in today's ASR engines, for example the Mel-Frequency Cepstral Coefficients (MFCC) or the Perceptual Linear Prediction (PLP) coefficients [31], are designed so that they imitate the human ear frequency resolution. Therefore including hearing properties in a noise reduction pre-processor can improve the robustness of those features. Furthermore, from a philosophical perspective, speech production and perception have evolved hand in hand over the centuries from the dawn of mankind. Hence, it seems sensible to argue that any sound produced by humans is only important for intelligibility if it can be perceived by the human ear. Otherwise, that sound would have never been developed. To make this idea clearer, it can be noted that an infant, while learning how to "produce" speech, can only repeat what he can hear or "perceive" from the sounds produced around him, particularly from his parents. For this reason it can be concluded that any sound which is not perceived by the human ear is redundant for the meaning the speaker seeks to convey. Therefore, even for a machine, it should be advantageous to decode an input speech signal which has been enhanced based on masking considerations. Actually this might explain the success of MFCC and PLP which made them the most popular features in the ASR industry.

Another class for speech enhancement methods is that of multi-microphone methods. To improve the overall performance, these methods take advantage of added degrees of

freedom achieved via the use of more than one microphone for sound acquisition. Different criteria have been proposed for the design of microphone arrays depending on the intended application. For example, for applications in a diffuse noise field, adaptive postfiltering is found to be appropriate [170, 117]. For a directional noise source, however, adaptive beamforming has been proposed [55, 92, 91].

Another method, which can also be viewed as a form of adaptive beamforming is the classical *adaptive noise canceling* (ANC) method [164]. ANC attempts to remove noise components which are correlated with the output of a second reference microphone in which the desired speech signal is known to be absent.

1.3 Research objectives

Most of the noise reduction methods used in practice are based on conventional frequency domain approaches, namely: spectral subtraction and Wiener filtering. The popularity of these methods is mainly due to their low computational load and ease of implementation. However, the emergence of new applications and the need for more robust performance of speech communication systems under noisy conditions, require the investigation of new approaches and the exploration of different tracks for speech enhancement research.

One such promising technique is the signal subspace approach (SSA) which is considered to be a powerful tool in various signal processing applications. For example, in array processing, the popular MUSIC algorithm, which is a signal subspace based technique, has been a considerable breakthrough in direction of arrival (DOA) estimation research [134, 135]. In speech enhancement, however, the use of SSA remains rather modest. After its first introduction in 1991 by Dendrinos *et al* [32], using singular value decomposition (SVD), and later in 1993 by Ephraim and Van Trees [40, 41], using eigenvalue decomposition (EVD), the SSA failed to attract much of the researchers' attention. The work performed so far has mainly focused on extending the white noise assumption of the original method, to the more practical colored noise case. Other researchers have attempted to reduce the relatively costly computational load of the SSA.

This latter problem is the main reason that hindered the use of the SSA in practice and discouraged researchers (especially from the industry) to pursue any further work in this direction. Fortunately, the silicon technology is rapidly developing and faster digital signal processors (DSP's) are continuously put to the market. Therefore the computational load

issue is becoming less significant compared to the gain in performance that can be achieved by employing the SSA instead of the conventional frequency domain methods.

Motivated by the rapid advances in DSP technology, and by the will to provide novel techniques to speech enhancement research, we seek in this thesis to improve the performance of SSA by further analyzing it, examining its drawbacks and uncovering its unexploited capabilities. Our interest in the SSA is particularly stirred by its enormous success in other applications, and also by the fact that the basic signal subspace method has shown a considerably better performance than, for instance, basic spectral subtraction [41].

During the research conducted for this thesis, the achieved results have intensified our belief that paying more attention to the SSA may pave the road to a possible breakthrough in speech enhancement research, thus opening new horizons and allowing for new applications that may have been previously considered to be infeasible.

1.4 Main contributions

In this thesis, we are interested in the signal subspace approach to reduce additive broadband noise such as car noise, jet cockpit noise or air conditioning noise for speech communications applications. Our main contributions can be summarized as follows:

Auditory Masking

Recently, new methods exploiting the human auditory masking properties have been successfully employed to improve the performance of the frequency domain suppression filters. However, since the available masking models are usually developed in the frequency domain, it is not clear how they can be applied in the SSA.

In this thesis we present and investigate a Frequency to Eigendomain Transformation (FET) which permits to calculate a perceptually based eigenfilter. This filter yields an improved result where better shaping of the residual noise, from a perceptual perspective, is achieved. The proposed method can also be used in the general case of colored noise.

We note that by itself, the FET is not a new mathematical concept. The novelty here is in bringing together two previously known relationships from signal processing and using them in the context of the signal subspace approach for speech enhancement, allowing the incorporation of the masking properties of the human auditory system. The developed

method is referred to as the Perceptual Signal Subspace (PSS) method[2].

The performance of this method has been evaluated by informal listening tests, spectrogram illustrations and subjective listening tests. These experiments have revealed the benefit of the proposed approach which results in a less annoying musical noise compared to other SSA methods. Actually, the word musical noise may no longer be suitable to describe the residual noise in this case. Indeed, for a given speech signal, the spectrum of the residual noise is shaped by PSS in such a way that its characteristics are found to be relatively similar regardless of the original background noise.

In addition to that, the FET has been used to analyze the SSA via a filterbank interpretation. This approach allows to understand the effects of SSA on the speech signal from a frequency domain perspective. Doing so, some phenomena related to SSA performance, reported in the literature, were explained leading to better design decisions for PSS.

Reducing the computational load

The main handicap of the SSA, as discussed earlier, is its relatively high computational load. In this thesis we provide a simple technique to reduce the complexity without introducing any additional distortion to the signal. This is achieved by reducing the rate at which the signal subspace filter is updated by exploiting the stationarity property of the speech signal within one frame of a specific length.

This technique, while preserving the same noise suppression performance, can considerably reduce the computational load. This result has been verified experimentally. This novel implementation technique is referred to as the Frame-Based EVD method (FBEVD).

Multi-microphone adaptive postfiltering

Our contribution to the signal subspace approach for speech enhancement is further extended to cover the multi-microphone case. We propose a generalization of the single microphone method into a multi-microphone design by applying the SSA to a composite input speech vector formed by samples from the different available microphone signals. Then, we exploit one property of the EVD of the covariance matrix of the extended problem, called

[2]The majority of the material related to PSS has been presented in ICASSP 2001 student forum [76], IWAENC 2001 [79], ICASSP 2002 [80] and as a journal paper in IEEE Trans. on Speech and Audio Processing [81].

the composite covariance matrix, to develop an improved technique with higher noise reduction capabilities while introducing insignificant signal distortion. This is achieved by performing averaging in the eigendomain to calculate the subspace filter coefficients. This method is called the Multi-microphone signal subspace method with Eigen-Domain Averaging (MEDA)[3]. By design, the MEDA can be transformed into an adaptive postfiltering technique which is experimentally found to be especially powerful in diffuse noise fields while being little sensitive to changes in the reverberation time.

Room response simulation

To evaluate the performance of the multi-microphone methods, we need to have a tool to digitally simulate the effects of room reverberation on speech signals. To this end, we provide a generalization of the popular image method [2], by allowing it to have frequency dependent reflection coefficients. This added design flexibility offers the possibility to simulate environments closer to real life conditions hence acquiring a more realistic judgment on the performance of the evaluated microphone array methods.

The proposed method is based on a subband scheme where the full band room impulse response is decomposed into several subband impulse responses. This design offers a straight forward fast implementation which permits to reduce the complexity hence saving the valuable simulations time. This method is called the subband Room Simulator (SRS)[4]

1.5 Thesis organization

This thesis is organized as follows.

Since an important part of this thesis consists of applying the masking properties to the signal subspace approach for speech enhancement, a brief description of the human hearing is presented in Chapter 2. Particularly, the phenomenon of masking is presented and some of the most popular masking models developed in the literature are described.

Chapter 3 consists of a review of some popular single microphone speech enhancement methods especially those which are closely related to the context of this thesis. The chapter includes an introductory section which covers some of the mathematical and signal pro-

[3]Parts of this method was presented in ICASSP 2001 [78].

[4]This method was presented in ICASSP 2000 [77].

cessing concepts important to this thesis. This section also serves as a reference for the notation and terminology used throughout the thesis.

The main method in the context of this thesis, namely the signal subspace approach (SSA), is thoroughly described in Section 3.4 where it is introduced and analyzed. The different linear estimators used to design the eigendomain filter are presented. At the end of the chapter the computational concerns about the method are raised and remedies proposed in the literature are given. Finally, the colored noise issue and the methods proposed to handle it are discussed.

In Chapter 4, we introduce the Frequency to Eigendomain Transformation (FET) and we describe how it is implemented as a matrix vector product making it suitable for implementation on digital computers. Using the FET, interpreted in a filterbank framework, we provide an analysis of the SSA which sheds more light on it from a frequency domain perspective. This analysis serves to better understand the advantages and shortcomings of the SSA related methods found in the literature.

The FET is used in Chapter 5 to design the novel PSS method. A full algorithm description is provided. Also in this chapter, we describe the novel Frame-Based EVD technique for a fast SSA implementation.

Chapter 6 is dedicated to the multi-microphone class for speech enhancement. The chapter begins with a literature review of the common noise field models and the most popular microphone array methods. After that, the novel MEDA is introduced and analyzed.

In Chapter 7 we present the novel subband room simulator. We describe the underlying algorithm and we quantify the achieved computational savings. At the end of this chapter we provide experimental results which verify the accuracy of this method as compared to the original image method.

Experimental results to assess the performance of the proposed novel speech enhancement methods, namely PSS, FBEVD and MEDA, are presented in Chapter 8. These results show the superiority of those methods over competing techniques.

Finally a conclusion and suggestions for future research are presented in Chapter 9.

1.6 Basic notation

Some of the basic notations and different mathematical symbols used in this thesis are as follows:

$\mathbf{x}$: Vectors are represented with small letters in bold. Any deviation from this notation will be properly clarified. Throughout the thesis, vectors are considered to be column vectors.

$\mathbf{A}$: Matrices are represented with capital letters in bold.

$(\cdot)^*$: Complex conjugate.

$(\cdot)^T$: Transpose operator.

$(\cdot)^H$: Hermitian operator.

$E\{\cdot\}$: The expected value operator.

$Re\{\cdot\}$: The real part of a complex number.

$Im\{\cdot\}$: The imaginary part of a complex number.

$\mathcal{R}\{\mathbf{A}\}$: Range or column space of matrix $\mathbf{A}$.

rank$\{\mathbf{A}\}$: The rank of matrix $\mathbf{A}$.

tr$\{\mathbf{A}\}$: The trace of matrix $\mathbf{A}$.

$\mathcal{F}\{\cdot\}$: The (discrete-time) Fourier transform.

$\mathcal{F}^{-1}\{\cdot\}$: The inverse Fourier transform.

Chapter 2

The Human Auditory Masking

One major contribution of this thesis, is the incorporation of the human auditory masking properties into the SSA. To be able to understand the benefit of this approach, it is important to first have a minimal understanding of the human hearing mechanism and the resulting masking phenomenon. To this end, we briefly discuss in this chapter the most important aspects of human hearing and the anatomy of the human auditory device, namely the ear. We next explain the masking properties and we provide a survey of the most popular masking models developed in the literature to mimic those properties. The interested reader can find more details about human hearing and psychoacoustics for example in [171] and [124].

2.1 An overview of hearing

Sound waves are captured by the ear, and converted into electric impulses transported by the auditory nerve to the part of the brain responsible for hearing. The ear, which constitutes the main part of the human hearing system, contains three parts: the outer ear, the middle ear and the inner ear, as shown in Figure 2.1.

The outer ear consists of the pinna, the ear canal and the eardrum. Sound pressure variations, transmitted via the ear canal, are converted into mechanical energy by inducing the vibration of the eardrum. This energy is then amplified in the middle ear through the vibration of the hammer (malleus), anvil (incus) and stirrup (stapes) along with the eardrum. The stirrup is connected to the oval window which is the entrance to the inner ear.

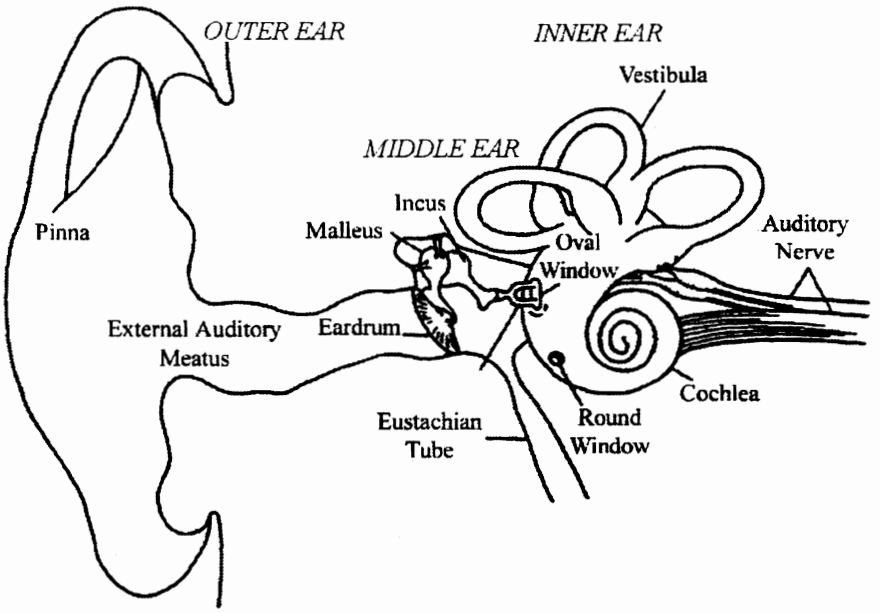


Fig. 2.1 Anatomy of the human ear [124].

The inner ear consisting of the cochlea is shaped like a snail, and is filled with fluids. The cochlea is divided along its length by two membranes. One of them, the basilar membrane (BM), supports the *organ of Corti* with its sensory hair cells, and plays an important role in hearing. This membrane detects the vibrations of the stirrup via the surrounding fluids and oscillates accordingly. The movements of the basilar membrane are sensed by the hair cells initiating the neural firing that lead to the perception of sound. A cross section of the cochlea is shown in Figure 2.2.

2.1.1 The basilar membrane

The basilar membrane is narrow and stiff at its base (near the end of the middle ear) and becomes wider and less stiff at its apex. The cochlea forms 2.5 turns allowing a BM length of about 32 mm [171]. Due to these physiological properties, each point on the BM is more sensitive to one distinct frequency called the *characteristic frequency*. Regions at the base are more sensitive to high frequencies whereas regions at the apex are more sensitive to low frequencies.

A sinusoidal sound traverses the basilar membrane as a traveling wave causing it to vibrate entirely at the frequency of the tone. The amplitude of the vibration, however, varies, being strongest at that point whose characteristic frequency matches the tone frequency [124]. The hair cells corresponding to the vibration peak detect this motion and fire accordingly allowing the brain (which gets this information via the auditory nerve) to identify the frequency and amplitude of the input signal.

The hair cells are approximately uniformly distributed along the BM whereas their characteristic frequencies have a logarithmic distribution. This phenomenon is the basis for the critical band analysis of the human auditory perception. A detailed picture of the basilar membrane and the organ of Corti is shown in Figure 2.2.

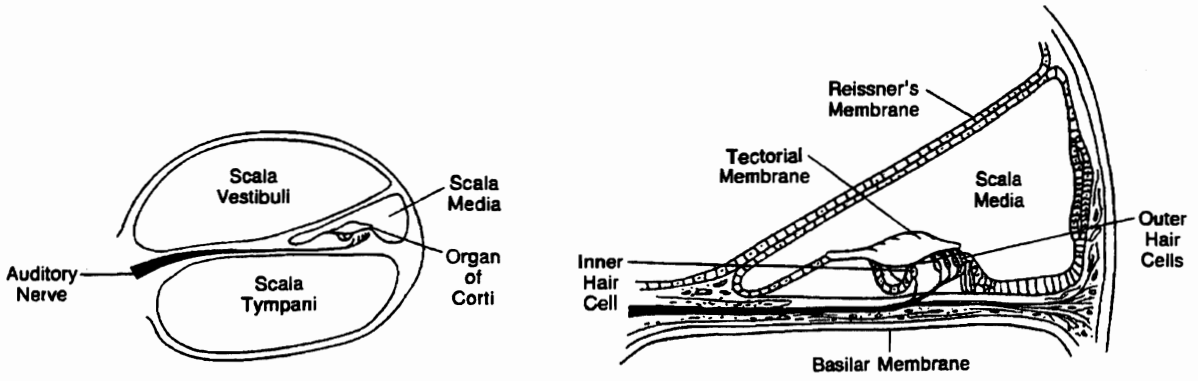


Fig. 2.2 A cross section of the cochlea. The organ of corti can also be seen in the righthand side figure [124].

2.1.2 Critical bands

A critical band (CB) is a range of frequencies the edges of which indicate an abrupt change in subjective responses [171]. Less technically, it represents a bandwidth within which the human ear ability to resolve different frequencies is diminished or almost impaired.

The bandwidth of the CB was first quantified experimentally by Fletcher [46, 171, 124]. In his experiment a tone is masked by a band of noise centered at the tone's frequency. The intensity of the tone was set so that it is inaudible in the presence of the noise. The bandwidth of the noise is then decreased gradually until the tone becomes audible again. The experiment is then repeated for different frequencies until all corresponding

Table 2.1 Critical band center and edge frequencies [88].

Band No.	Center Frequency (Hz)	Edge frequencies (Hz)
1	50	0-100
2	150	100-200
3	250	200-300
4	350	300-400
5	450	400-510
6	570	510-630
7	700	630-770
8	840	770-920
9	1000	920-1080
10	1170	1080-1270
11	1370	1270-1480
12	1600	1480-1720
13	1850	1720-2000
14	2150	2000-2320
15	2500	2320-2700
16	2900	2700-3150
17	3400	3150-3700
18	4000	3700-4400

such bandwidths have been quantified. As expected these bandwidths are found to increase logarithmically with frequency.

A perceptual measure, called the Bark scale, relates the acoustic frequency to the non-linear perceptual frequency resolution, in which one Bark covers one critical bandwidth. The analytical expression used to map the frequency f (in Hertz) to the critical-band rate z (in Barks) is [171]

$$z(f) = 13\arctan(0.00076f) + 35\arctan[(\frac{f}{7500})^2] \tag{2.1}$$

The bandwidth of each CB can be related to the center frequency as follows [171]

$$BW(f) = 25 + 75(1 + 1.4(f/1000)^2)^{0.69} \tag{2.2}$$

A simpler relationship is given by [171]

$$BW(f) = \begin{cases} 100 & f < 500 \\ 0.2f & f > 500 \end{cases} \tag{2.3}$$

which is more intuitive since it makes clear that the critical bandwidths are constant up to 500 Hz and increase with frequency thereafter.

Although equation (2.1) provides a continuous mapping from linear to bark scale many perceptually based speech processing algorithms use a quantized bark number to index the critical bands within the frequency range of interest. These bark indices together with their corresponding critical band center and edge frequencies are shown in Table 2.1.

2.1.3 The absolute threshold of hearing

Even under the most convenient conditions, the human ear still has its limits as to what extent it can detect sounds. This limit is quantified by the so-called absolute threshold of hearing. The absolute threshold of hearing (or the threshold in quiet) is the sound pressure level of a pure tone that is just audible in a noiseless environment. This threshold is well approximated by the following nonlinear function [148]

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5\exp(-0.6(f/1000 - 3.3)^2) + 10^{-3}(f/1000)^4 \quad \text{dB} \tag{2.4}$$

which is representative of a young listener with acute hearing.

2.2 Auditory masking

Due to the physiological properties of the human hearing system, weaker sounds are masked by stronger sounds taking place close in frequency or time. The reason for this phenomenon is that the activity caused by the weaker signal (the maskee) in the auditory system is not detected due to the activity caused by the stronger one (the masker).

Two types of masking can be recognized: *simultaneous and temporal masking.* Simultaneous masking, which is a frequency domain phenomenon, takes place when both the masker and the maskee are present at the same time, i.e. simultaneously. Temporal masking occurs when the maskee is presented to the ear after (forward masking) or before

(backward masking) the masker.

2.2.1 Temporal masking

There are two types of temporal masking. The first is the forward masking which takes place when the maskee occurs within 200 msec *after* the end of the masker. It is due to the fatigue or latency of the neurons. In fact the auditory neurons cannot fire again after a firing until a *latency period* of 1-3 msec have elapsed [124]. This suggests that the two sounds should also be close enough in frequency.

The second type is backward masking which occurs when the maskee comes within 20 msec *before* the masker and is due to an interference or blockage of the neural information on its way to the brain [124].

2.2.2 Simultaneous masking

Simultaneous masking, also referred to as frequency or spectral masking, occurs when both the masker and the maskee are presented to the ear at the same time. It is based on the frequency resolution of the basilar membrane, or critical band analysis. This type of the masking phenomenon can be explained as follows.

Consider a tone (the masker) at some frequency f_0. A second tone (the maskee) at frequency $f_0 + \delta f$ will be inaudible if its intensity is below some threshold. This masking threshold, which depends on δf, is found to be asymmetric in the sense that it is easier to mask tones at higher frequencies ($\delta f > 0$) than tones at lower frequencies ($\delta f < 0$). Besides, the slope of the masking curve was found to be dependent on the intensity of the masker at higher frequencies whereas at lower frequencies, the slope is constant [171].

Another observation is that the bandwidth of the masking curve increases as f_0 increases, that is a wider range of frequencies is affected by the masker if the frequency of the latter increases. This frequency range is nothing but the critical bands described earlier.

2.3 Masking Models

In order to exploit the masking properties described above in various speech and audio applications, different masking models emulating the ear's behaviour have been developed. Models for both simultaneous and temporal masking have been proposed with various

complexity and precision. However, while temporal masking was found to be useful, for example in wideband audio coding [82], it has not been as widely used, especially in speech enhancement applications, as simultaneous masking. This is because it was found to be less important than simultaneous masking, in addition to its being more difficult to quantify. For this reason we only focus in this thesis on models for simultaneous masking.

Most of the masking models developed follow almost the same basic steps: critical band analysis, application of a spreading function to take care of inter-critical band masking, subtraction of the masking offset depending on the tonality of the masker and finally comparison with the absolute threshold of hearing. To facilitate the readers understanding of these steps, we give in section 2.3.1 a detailed description of one of the first developed models, namely Johnston's model [88]. This model is mainly based on the work done by Schroeder [136] *et al.* and has been designed for audio coding applications. The reason we chose to describe this model in detail is its simplicity which makes the different calculation steps easier to understand. In Section 2.3.2 we describe the MPEG models which are also used in audio coding. A special emphasis is put on the MPEG model 1 since it is the masking model opted for in this thesis. In section 2.3.3, a survey of other masking models will be given.

2.3.1 Johnston's model

Johnston's model can be considered to be less sophisticated than other models available nowadays, especially in the context of audio coding. In speech enhancement, however, it still can provide satisfactory results and is the model used for example in [157, 150]. In this model, the effect of individual masking components on the global masking threshold is additive. Besides, although equation (2.1) is continuous, this model uses discrete Bark numbers corresponding roughly to the upper band edges of the critical bands of interest, whose number is B. In audio applications, 24 critical bands that cover the human hearing range, are typically used. For a speech signal sampled at 8 KHz on the other hand, just 18 critical bands are retained. The center frequencies of the CB's as well as their upper and lower edge frequencies are shown in Table 2.1.

Johnston's model consists of the following steps:

1. For every analysis frame, spectral energies within every critical band is summed to

obtain a unique energy $E(i)$ with i being the bark number,

$$E(i) = \sum_{k=k_{li}}^{k_{hi}} |X(k)|^2 \quad i = 1, \ldots, B \tag{2.5}$$

where k_{li} and k_{hi} are the lower and upper limits of critical band i respectively, and $X(k)$ is the DFT of the current speech frame. This step accounts for the critical band analysis of the human ear where all tones with frequencies within the i^{th} critical band are represented with a single tone with energy $E(i)$.

2. Inter-band masking is accounted for by convolution with a spreading function. This function has lower and upper skirts of +25 dB and -10 dB per critical band respectively and is given by [136]

$$SF_{(dB)}(z) = 15.81 + 7.5(z + 0.474) - 17.5\sqrt{1 + (z + 0.474)^2} \quad (dB) \tag{2.6}$$

The spread bark spectrum is then obtained as follows[1]

$$C(i) = \sum_{j=1}^{B} E(j)SF(i - j) \tag{2.7}$$

3. Next, the masking threshold is obtained by subtracting a relative threshold offset depending on the masker type, tone-like or noise-like. The tonality α is measured using the spectral flatness measure (SFM) in dB:

$$\text{SFM} = 10 \log_{10} \frac{G}{A} \tag{2.8}$$

where G is the geometric mean of the signal's power spectrum and A is its arithmetic mean. The tonality α is then calculated as

$$\alpha = \min \left\{ \frac{\text{SFM}}{\text{SFM}_{\text{max}}}, 1 \right\} \tag{2.9}$$

$SFM_{\text{max}} = -60$ is defined as the SFM of a sine wave. The relative offset is then

[1]Note that to correctly perform the convolution the spreading function should be converted from its decibel representation to a linear scale.

calculated as follows

$$O(i) = \alpha(14.5 + i) + (1 - \alpha)5.5 \tag{2.10}$$

In [157], the above approach is considered to be complicated and a simpler approximation is used instead to calculate the tonality. It is based on the observation that speech is typically tone like ($\alpha = 1$) in low frequencies and noise-like ($\alpha = 0$) in high frequencies [139].

The masking threshold is then calculated as follows

$$T(i) = C(i)10^{-O(i)/10} \tag{2.11}$$

4. After comparing it with the absolute threshold of hearing and retaining the maximum of the two, the masking threshold is mapped from the bark scale to the linear frequency scale.

2.3.2 The MPEG models

The MPEG standard provides two psychoacoustic models for use in the audio coder [72, 12]. These models calculate the signal-to-mask ratio (SMR) which is the difference (in dB) between the maximum signal level and the minimum masked threshold level.

The psychoacoustic model 1 [12] is designed for use with Layers 1 and 2 of the MPEG audio standard whereas the model 2, which is more sophisticated, is mainly designed for use with the Layer 3.

Psychoacoustic model 1

This model calculates the magnitude spectrum of the input signal by the FFT. Tonal and nontonal (noise-like) components of the spectrum are then identified. The masking threshold of each of these individual components is calculated and the resulting individual thresholds are summed linearly to obtain the global masking threshold. A masking component at a particular frequency is discarded if it is below the absolute threshold of hearing at that frequency.

The masking threshold of a tonal component is given by

$$T_{tm}(j,i) = X_{tm}(j) + O_{tm}(j) + SF(j,i), \tag{2.12}$$

where $T_{tm}(j,i)$ is the masking threshold at i barks due to the masking component located at j barks. $X_{tm}(j)$ is the sound pressure level (in dB) of the masking component with critical band index j. The function $O_{tm}(j)$ is the threshold offset given by

$$O_{tm}(j) = -1.525 - 0.275j - 4.5 \text{ dB} \tag{2.13}$$

Similarly, the masking threshold of each nontonal component is given by

$$T_{nm}(j,i) = X_{nm}(j) + O_{nm}(j) + SF(j,i), \tag{2.14}$$

where

$$O_{nm}(j) = -1.525 - 0.175j - 0.5 \text{ dB} \tag{2.15}$$

$SF(j,i)$ is the spreading function given by

$$SF(j,i) = \begin{cases} 17(dz+1) - 0.4X(j) - 6\text{dB} & -3 \le dz < -1 \text{ bark} \\ (0.4X(j)+6)dz\text{dB} & -1 \le dz < -0 \text{ barks} \\ -17dz\text{dB} & -0 \le dz < -1 \text{ bark} \\ -(dz-1)(17-0.15X(j)) - 17\text{dB} & -1 \le dz < -8 \text{ barks} \end{cases} \tag{2.16}$$

where $dz = i - j$ in barks. The spreading function has no effect on regions of the spectrum that are outside the range of -3 to 8 barks on the critical band rate scale, relative to the location of the masking component. $X(j)$ in (2.16) stands for either $X_{tm}(j)$ or $X_{nm}(j)$.

Further implementation details of this model will be described in Section 5.4.

Psychoacoustic model 2

In the psychoacoustic model 2, the maskers are not isolated and classified as tonal or non-tonal. Instead, a tonality index (similar to Johnston's model) is calculated based on an *unpredictability measure*. This measure is used to interpolate between masking thresholds produced by the extreme cases of noise-masking-tone (NMT) and tone-masking-noise

(TMN). The unpredictability measure is obtained by comparing the actual magnitude and phase spectral values of the current frame with values that were extrapolated from the previous two frames. This measure will be close to zero for a tonal signal and close to one for a noise-like signal. This approach has been found by Johnston to be more appropriate than the one based on the spectral flatness measure (SFM) [87].

The offset value is then calculated as an interpolation between the offset for NMT (a constant 5.5 dB) and for TMN (which varies with frequency from 24.5 dB to about 40 dB).

Another model has also been provided by MPEG for use in the Advanced Audio Coding (AAC) standard [73]. The AAC model is very similar to the MPEG model 2 with the difference mainly in the offset value for the TMN which is in this case, 18 dB for all bands.

2.3.3 Other masking models

In addition to Johnston's model and the MPEG models, there exist many other models which were developed for various speech and audio applications. For example Colomes *et al.* [23] proposed a model for use in an audio codec (as well as in an objective perceptual meter). This model emulates the ear canal effects using a pre-filter then applies FFT to obtain 2048 spectral lines grouped into 600 bands equally spaced on the bark scale.

Some other models have been developed to be used in objective audio/speech quality measurements. In this framework, we can mention the PEAQ (for audio) [74, 149] and the PESQ (for speech) [75] models proposed by the International Telecommunication Union. Beerends *et al.* [6] also proposed a model which differs from that of Johnston in that no tonality of the signal is calculated and the difference in masking between tonal and non-tonal components is accounted for by a compressed loudness measure. In addition to that, a more complex spreading function with level dependent upper slope, is used. Other models used for evaluation of audio quality can be found in [125] and in [67] where a filterbank is used for the frequency transformation instead of the Fourier transform. Filter bank analysis is also used in general purpose models such as those described in [27, 28, 131].

Chapter 3

Single Channel Speech Enhancement: Background Material

Most speech communication devices use a single microphone for signal acquisition mainly to reduce the overall cost and to comply with the spatial constraints imposed by the application. For this reason, a variety of single channel speech enhancement methods and techniques have been developed during the past decades and are widely used in practice. In this chapter we go through the most popular methods by studying their underlying theory, their advantages and drawbacks.

To this end, we first start by introducing some mathematical tools important to understand the material presented in this chapter and in the thesis in general. Then, we discuss the widely used frequency domain methods with special emphasis placed on those methods which exploit the masking properties of the human ear. The interested reader can find more details for example in [31] and in the references we cite here. This thesis is mainly concerned about the signal subspace approach, that is why a thorough detailed discussion of that technique is presented. Finally we attract the reader's attention to the crucial problem of noise estimation by providing a brief survey of some of the proposed techniques in the literature.

3.1 Mathematical background

In this section we briefly review some of the useful definitions and properties of digital signal processing, linear algebra and random processes which are relevant to the understanding of the speech enhancement concepts described in this thesis. The reader can find additional details for example in [48, 65, 64, 123].

3.1.1 Linear Algebra

Consider an $n \times m$ matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m]$, where $\mathbf{a}_i$'s are n dimensional column vectors. The range of $\mathbf{A}$ is defined as

$$\mathcal{R}\{\mathbf{A}\} = \{\mathbf{x} \in \mathbb{C}^n : \exists\, \mathbf{y} \in \mathbb{C}^m, \mathbf{A}\mathbf{y} = \mathbf{x}\} \tag{3.1}$$

That is, it is the subspace spanned by the columns of $\mathbf{A}$, hence the alternative name, column span of $\mathbf{A}$. The dimension of $\mathcal{R}\{\mathbf{A}\}$ is given by the rank of $\mathbf{A}$ which is the number of its linearly independent column vectors.

A matrix $\mathbf{A}$ is said to be positive semidefinite, $\mathbf{A} \geq 0$, if for all non-zero vectors $\mathbf{x}$, $\mathbf{x}^H\mathbf{A}\mathbf{x} \geq 0$. $\mathbf{A}$ is said to be positive definite, $\mathbf{A} > 0$, if $\mathbf{x}^H\mathbf{A}\mathbf{x} > 0$.

A non-zero vector $\mathbf{u}$ is an eigenvector of a $n \times n$ matrix $\mathbf{A}$ if it satisfies $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$, where λ is the corresponding eigenvalue. The eigenvalues and eigenvectors satisfy the following properties

1. The non-zero eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ corresponding to distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ are linearly independent.

2. A Hermitian matrix $\mathbf{A}$, that is $\mathbf{A}^H = \mathbf{A}$, has real eigenvalues and is positive definite ($\mathbf{A} > 0$) if and only if these eigenvalues are strictly positive.

3. The eigenvectors of a Hermitian matrix, corresponding to distinct eigenvalues are orthogonal. i.e. $\mathbf{u}_i^H\mathbf{u}_j = 0$ for $\lambda_i \neq \lambda_j$.

4. Any Hermitian matrix $\mathbf{A}$ may be decomposed as

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H = \lambda_1\mathbf{u}_1\mathbf{u}_1^H + \lambda_2\mathbf{u}_2\mathbf{u}_2^H, \ldots + \lambda_n\mathbf{u}_n\mathbf{u}_n^H \tag{3.2}$$

where λ_i are the eigenvalues of $\mathbf{A}$ and $\mathbf{u}_i$ are a set of orthonormal eigenvectors. Here $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ is the eigenvalue matrix and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ is the eigenvector matrix. Note that $\mathbf{U}$ is unitary[1], that is $\mathbf{U}^H\mathbf{U} = \mathbf{I}$. We refer to (3.2) as the eigenvalue decomposition (EVD) of matrix $\mathbf{A}$.

3.1.2 Discrete-time Signal Processing

Consider an analog continuous signal $x_c(t)$ with bandwidth $F_s/2$. The discrete representation $x(n)$ of this signal is obtained by sampling $x_c(t)$ according to the *Nyquist criterion*,

$$x(n) = x_c(nT_s) = x_c(t)|_{t=nT_s} \tag{3.3}$$

where $T_s = 1/F_s$ and F_s is the sampling frequency.

The Discrete-Time Fourier Transform (DTFT) of $x(n)$ is the complex-valued function of the continuous variable ω (angular frequency) defined by

$$\mathcal{F}\{x(n)\} = X(\omega) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} \tag{3.4}$$

The Inverse Discrete-Time Fourier Transform (IDTFT) is given by

$$\mathcal{F}^{-1}\{X(\omega)\} = x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\omega)e^{j\omega n}d\omega \tag{3.5}$$

Throughout the thesis, time domain waveforms are represented by small letters and the corresponding capital letters should be understood as their Fourier Transforms.

The DTFT, a complex quantity in general, can be written as

$$X(\omega) = |X(\omega)| \cdot e^{j\angle X(\omega)} \tag{3.6}$$

where $|X(\omega)|$ and $\angle X(\omega)$ are the amplitude and the phase of $X(\omega)$ respectively.

While the DTFT remains a useful mathematical tool, a more practical representation of the frequency domain is achieved via the *Discrete Fourier Transform* (DFT) which is a

[1]Note that for real entries, Unitary matrices are simply called orthogonal matrices. In addition, Hermitian matrices are then symmetric matrices, and the Hermitian operator $(.)^H$ is replaced by a transposition operator $(.)^T$.

function of an integer variable k. For a finite-length sequence $x(n)$ defined over the interval $[0, N-1]$, the N-point DFT of $x(n)$ is

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N} \quad \text{for } k = 0, \ldots, N-1 \tag{3.7}$$

If $x(n) = 0$ outside the interval $[0, N-1]$, its DFT is equal to the DTFT sampled at N equally spaced frequencies in $[0, 2\pi]$, that is

$$X(k) = X(\omega)|_{\omega = 2\pi k/N} \tag{3.8}$$

The N-point DFT will be represented using the vector

$$\mathbf{X} = [X(0), \ldots, X(N-1)]^T \tag{3.9}$$

The vector $\mathbf{X}$ will then be related to the DTFT $X(\omega)$ according to (3.9) and (3.8).

In practice the DFT is efficiently calculated using the Fast Fourier Transform (FFT) which reduces the number of required complex multiplications, when N is a power of 2, from N^2 to $\frac{N}{2}\log_2 N$.

3.1.3 Stochastic processes

Speech may be considered as a deterministic signal if a specific waveform is to be processed or analyzed. However, it may be also viewed as a random process if one is considering the ensemble of all possible waveforms in order to design a system that will optimally process the speech signal. In this thesis the speech signal is considered to be a random signal.

Within a short observation interval of about 20-40 msec, a speech signal $x(n)$ is considered to be a realization of a zero mean and *wide-sense stationary* (WSS) random process with autocorrelation function

$$\tilde{r}_x(p) = E\{x(n)x^*(n+p)\} \tag{3.10}$$

which is a conjugate symmetric function of p, i.e. $\tilde{r}_x(p) = \tilde{r}_x^*(-p)$.

The autocorrelation sequence is often represented in matrix form. Consider the vector

$\mathbf{x} = [x(n), x(n-1), \ldots, x(n-P+1)]^T$, the covariance matrix is defined as follows

$$\tilde{\mathbf{R}}_x = E\{\mathbf{x}\mathbf{x}^H\} = \begin{bmatrix} \tilde{r}_x(0) & \tilde{r}_x^*(1) & \cdots & \tilde{r}_x^*(P-1) \\ \tilde{r}_x(1) & \tilde{r}_x(0) & \cdots & \tilde{r}_x^*(P-2) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{r}_x(P-1) & \tilde{r}_x(P-2) & \cdots & \tilde{r}_x(0) \end{bmatrix} \tag{3.11}$$

It can be immediately observed that $\tilde{\mathbf{R}}_x$ is a non-negative definite Hermitian Toeplitz matrix, ($\tilde{\mathbf{R}}_x = \text{Toeplitz}\{\tilde{r}_x(0), \tilde{r}_x(1), \ldots, \tilde{r}_x(P-1)\}$), hence all its eigenvalues are real valued and non-negative.

The power spectral density (PSD) is defined as the DTFT of the autocorrelation function,

$$\tilde{\Phi}(\omega) = \sum_{p=-\infty}^{\infty} \tilde{r}_x(p) e^{-jp\omega} \tag{3.12}$$

The inverse relationship is obtained via the IDTFT defined in (3.5). We note that the power spectrum is real and nonnegative, i.e. $\tilde{\Phi}(\omega) \geq 0$ for all ω.

The PSD is also a theoretical tool and in practice it needs to be estimated from the observed data. Several methods for power spectrum estimation were developed and can be found in classical books like [94] or [116]. In this thesis we are especially interested in two of these methods namely: the periodogram and the Blackman-Tukey estimators.

The periodogram is defined as follows

$$\Phi_{per}(\omega) = \sum_{p=-N+1}^{N-1} r_x(p) e^{-jp\omega} \tag{3.13}$$

where $r_x(p)$ is the biased autocorrelation estimate calculated from one finite realization $x(n)$ of the random process as follows

$$r_x(p) = \frac{1}{N} \sum_{n=0}^{N-|p|-1} x(n)x^*(n+p), \quad p = -N+1, \ldots, N-1 \tag{3.14}$$

where it is assumed that $x(n) = 0$ for $n < 0$ and $n \geq N$. It can be verified that the

periodogram is directly related to the data as follows

$$\Phi_{per}(\omega) = \frac{1}{N}|X(\omega)|^2 \tag{3.15}$$

where $X(\omega)$ is the DTFT of $x(n)$ over $0 \leq n \leq N-1$.

The Blackman-Tukey estimator (BT), on the other hand, is defined as

$$\Phi_B(\omega) = \sum_{p=-P+1}^{P-1} r_x(p) w_b(p) e^{-j\omega p} \tag{3.16}$$

where $w_b(p)$ is a symmetric data window of length $2P-1$.

The properties of these two estimators will be further addressed in Section 4.4. Note that unless otherwise mentioned all PSD illustrations in this thesis are obtained via the BT estimator with a triangular (Bartlett) window with $P = 32$.

3.2 Frequency domain speech enhancement methods

Frequency domain methods for speech enhancement are widely used due to there simplicity. These methods mainly include spectral subtraction, Wiener filtering and their variants. In this section we review these methods and discuss some of their limitations and the cures available in the literature.

3.2.1 Spectral subtraction

In the spectral subtraction method, and its variants [7, 9, 10, 38, 107, 109, 119, 138, 155], noise reduction is performed in the frequency domain using a data independent transform, namely, the Discrete Time Fourier Transform (DTFT).

Let $x(n) = s(n) + w(n)$, be a noisy speech signal where $s(n)$ is the clean signal and $w(n)$ is an uncorrelated additive background noise. In the frequency domain we have

$$X(\omega) = S(\omega) + W(\omega) \tag{3.17}$$

In the particular case of power spectral subtraction, an estimate of the squared magni-

tude spectrum of the clean speech signal is computed as follows

$$|\hat{S}(\omega)|^2 = \max(|X(\omega)|^2 - |\hat{W}(\omega)|^2, 0) \tag{3.18}$$

where $|\hat{W}(\omega)|^2$ is the noise power spectrum estimate obtained during non speech activity periods. It is known that the human auditory system is relatively insensitive to phase distortion [160]. Hence the phase of the noisy signal is used to recover the time domain waveform as follows

$$\hat{s}(n) = \mathcal{F}^{-1}\{|\hat{S}(\omega)| \cdot e^{j\angle X(\omega)}\} \tag{3.19}$$

More generally, the spectral subtraction method can be formulated as follows [157]

$$|\hat{S}(\omega)| = H(\omega) \cdot |X(\omega)| \tag{3.20}$$

where $H(\omega)$ is an attenuation function given by [157]

$$H(\omega) = \begin{cases} \left(1 - \alpha \cdot \left[\frac{|\hat{W}(\omega)|}{|X(\omega)|}\right]^{\gamma}\right)^{\frac{1}{\gamma}}, & \text{if } \left[\frac{|\hat{W}(\omega)|}{|X(\omega)|}\right]^{\gamma} < \frac{1}{\alpha+\beta} \\ \left(\beta \cdot \left[\frac{|\hat{W}(\omega)|}{|X(\omega)|}\right]^{\gamma}\right)^{\frac{1}{\gamma}}, & \text{otherwise} \end{cases} \tag{3.21}$$

where α is an oversubtraction factor that controls the trade off between the level of the residual noise and the signal distortion, β is a spectral flooring parameter which adds a background that helps to mask the (usually annoying) residual noise. The exponent γ determines the sharpness/smoothness of the attenuation function. Common values are usually $\gamma = 2$ for power spectral subtraction and $\gamma = 1$ for amplitude spectral subtraction.

The attenuation function (or filter) $H(\omega)$ can be written (for the basic power spectral subtraction[2]) as a function of the instantaneous SNR defined as

$$\text{SNR}(\omega) = \frac{|X(\omega)|^2}{|\hat{W}(\omega)|^2} \tag{3.22}$$

in the following way

$$H(\omega) = \left[1 + \frac{1}{\text{SNR}(\omega)}\right]^{-1/2} \tag{3.23}$$

[2]That is equation (3.18), or equation (3.21) with $\alpha = 1$, $\beta = 0$ and $\gamma = 2$.

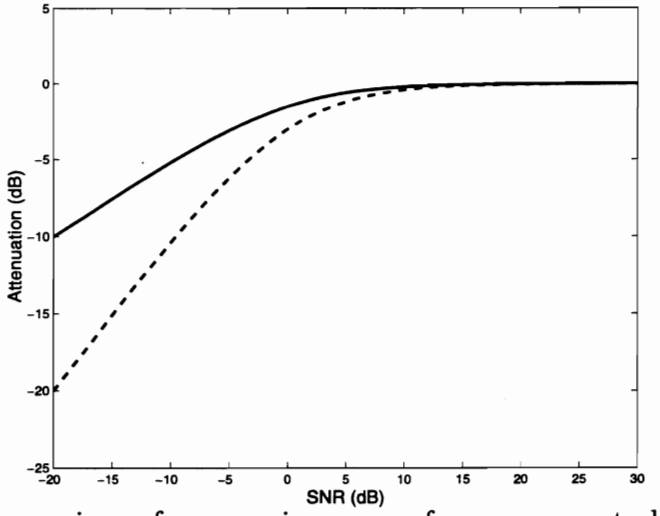


Fig. 3.1 Comparison of suppression curves for power spectral subtraction (continuous) and Wiener filtering (dashed) as a function of the signal to noise ratio.

A plot of this attenuation function is shown in Figure 3.1.

3.2.2 Wiener filtering

Another closely related method for noise reduction is Wiener filtering. In this framework, we consider $x(n)$, $s(n)$ and $w(n)$ to be the underlying random processes of the noisy signal, speech signal and noise signal respectively. The goal is to design a linear filter with $x(n)$ as input and outputs an estimate for $s(n)$, say $\hat{s}(n)$, which is optimal in the Minimum Mean Squared Error (MMSE) sense. This filter is obtained such that the error $E\{|s(n)-\hat{s}(n)|^2\}$, is minimized. In the frequency domain, and assuming that the clean speech and noise are uncorrelated, the solution is given by the classical Wiener filter,

$$H(\omega) = \frac{\tilde{\Phi}_s(\omega)}{\tilde{\Phi}_s(\omega) + \tilde{\Phi}_w(\omega)} \tag{3.24}$$

where $\tilde{\Phi}_s(\omega)$ and $\tilde{\Phi}_w(\omega)$ are the clean speech and noise PSD's respectively.

Redefining the instantaneous SNR now as,

$$\mathrm{SNR}(\omega) = \tilde{\Phi}_s(\omega)/\tilde{\Phi}_w(\omega),$$

the Wiener filter can be written as

$$H(\omega) = \left[1 + \frac{1}{\text{SNR}(\omega)}\right]^{-1} \tag{3.25}$$

A comparison of the suppression curves of spectral subtraction and Wiener filtering is shown in Figure 3.1. It can be seen that for low SNR values, Wiener filtering has a stronger attenuation than spectral subtraction.

3.2.3 Main limitation

Due to poor estimation of signal and/or noise statistics, both Wiener filtering and spectral subtraction suffer from a residual noise which has an annoying noticeable tonal characteristics [155]. This processing artifact, usually referred to in the literature as *musical noise*, results from spectral peaks randomly distributed over time and frequency. These peaks are usually attributed to the fluctuations in the suppression filter coefficients both over time and frequency.

Many solutions have been proposed to overcome this problem: averaging of magnitude spectra of adjacent frames [9], over-subtraction of noise and introduction of a spectral floor [7], soft-decision noise suppression filtering [119] and optimal MMSE estimation of the short-time spectral amplitude [38]. A non-linear spectral subtraction method was also proposed in which the subtraction factor depends non-linearly on a frequency dependent SNR [112]. The use of human ear masking properties is another approach proposed in the literature. Masking will be described in more details in the next section.

Despite this variety of techniques developed over the years, musical noise remains the major drawback of these frequency domain subtractive methods and further research is still needed to overcome this difficulty.

3.3 Speech enhancement based on auditory masking

An alternative promising solution to reduce the intensity of the musical noise, is to exploit the properties of the human auditory system [5, 25, 56, 150, 157]. The idea is based on the fact that a signal occurring close in time or in frequency to a stronger signal will be masked, that is, it will not be perceived by the human listener [171]. Following this principle, these methods attempt not to cancel a noise component as long as its presence is not perceived

because it is masked by a nearby speech component. Ideally, such processing would result in giving the spectrum of the residual noise a shape which closely follows that of the desired speech signal. This objective, which we will be referring to as noise shaping, is expected to eventually mask the residual noise.

While the use of auditory masking has been first exploited in audio coding [12, 88], it is also gaining popularity in other fields such as objective evaluation of audio/speech quality [74, 75, 149], noise reduction, and more recently echo cancellation [57]. In the following, we present three of the most popular noise reduction methods based on masking threshold. Other methods have also been introduced in the literature and the interested reader can find more details in the above mentioned references and the citations therein.

3.3.1 Virag's method

In Virag's method [157], spectral subtraction is implemented using the general filter form given in (3.21). Virag updates the oversubtraction and spectral flooring parameters α and β, respectively, according to the masking threshold calculated from the currently processed noisy frame. The dependency of these parameters on the masking threshold is based on the following relationships

$$\alpha_i(\omega) = F_\alpha[\alpha_{\min}, \alpha_{\max}, T(\omega)] \tag{3.26}$$

$$\beta_i(\omega) = F_\beta[\beta_{\min}, \beta_{\max}, T(\omega)] \tag{3.27}$$

where i is the current frame index. $T(\omega)$ is the masking threshold calculated from an initial estimate of the clean speech spectrum, obtained using the conventional spectral subtraction applied to the current frame. Virag uses Johnston's model [88] to calculate the masking threshold.

The parameters $\alpha_{\min}$, $\beta_{\min}$ and $\alpha_{\max}$, $\beta_{\max}$ are the minimal and maximal values of the filter parameters. The function (3.26) that controls the filter parameters is given by

$$F_\alpha[\alpha_{\min}, \alpha_{\max}, T(\omega)] = \alpha_{\max} \quad \text{if} \quad T(\omega) = \min[T(\omega)] \tag{3.28}$$

$$F_\alpha[\alpha_{\min}, \alpha_{\max}, T(\omega)] = \alpha_{\min} \quad \text{if} \quad T(\omega) = \max[T(\omega)] \tag{3.29}$$

Between these extreme values, interpolation based on $T(\omega)$ is employed. For example a linear interpolation can be used.

The idea is to exert more severe suppression when the value of the masking threshold is low, and a less severe suppression when the threshold is high. In the latter case, for instance, more noise is allowed to remain in the enhanced signal since the speech signal is expected to mask it. A similar approach is used for $F_\beta[\beta_{\min}, \beta_{\max}, T(\omega)]$ in (3.27).

To obtain a good tradeoff between residual noise and signal distortion, the following values have been proposed by Virag:

$$\alpha_{\min} = 1 \quad \text{and} \quad \alpha_{\max} = 6$$
$$\beta_{\min} = 0 \quad \text{and} \quad \beta_{\max} = 0.02$$

This method has been reported to provide a better noise suppression performance. Objective and subjective measures were provided for evaluation. Virag also found that the use of masking properties improves speech recognition accuracy [157].

3.3.2 Tsoukalas's method

In the method proposed by Tsoukalas *et al.* [150], a noise suppression filter is designed based on a psychoacoustically derived quantity of audible noise spectrum. The audible noise spectrum $A_w(\omega)$ is defined as follows,

$$A_w(\omega) = \begin{cases} \Phi_x(\omega) - \hat{\Phi}_s(\omega) & \text{if } \hat{\Phi}_s(\omega) \geq T(\omega) \\ \Phi_x(\omega) - T(\omega) & \text{if } \hat{\Phi}_s(\omega) < T(\omega) \end{cases} \tag{3.30}$$

where $T(\omega)$ is a masking threshold obtained from $\hat{\Phi}_s(\omega)$ using Johnston's model [88], $\hat{\Phi}_s(\omega)$ is a rough preliminary estimate of the clean speech PSD obtained for example using power spectral subtraction and $\Phi_x(\omega)$ is the noisy speech spectrum.

The idea here is to design a filter which would just suppress the audible noise as defined in (3.30) while keeping all other noise components as long as they are not audible. Such a filter is given by the following expression

$$H(\omega) = \frac{\Phi_x(\omega)}{a(\omega) + \Phi_x(\omega)} \tag{3.31}$$

where $a(\omega)$ is a threshold below which all frequency components are highly suppressed. Within one critical band with index i, $a(\omega)$ takes a unique value according to the following

relationship

$$a(\omega) = \max_{\omega_l(i) \leq \omega \leq \omega_h(i)} \left\{ \Phi_x(\omega) \left[\frac{\Phi_x(\omega)}{\max[\hat{\Phi}_s(\omega), T(\omega)]} - 1 \right] \right\}, \quad \omega_l(i) \leq \omega \leq \omega_h(i). \tag{3.32}$$

where $\omega_l(i)$ and $\omega_h(i)$ are the lower and upper boundaries of the critical band with index i, respectively.

For best performance, (3.32) requires an accurate estimate of the clean speech spectrum. Therefore, it was suggested in [150] to just rely on a single value of $\hat{\Phi}_s(\omega)$ per critical band in order to minimize the dependency on such estimates. This technique is referred to as *sparse speech estimation*. Note also that both the masking threshold $T(\omega)$ and the noise spectrum $\Phi_w(\omega)$ are assumed to be constant within one critical band.

Accordingly, the following alternative for (3.32) is proposed

$$a(\omega) = \frac{\Phi_w(i)}{\hat{\Phi}_{s,\min}(i)} [\Phi_w(i) + \hat{\Phi}_{s,\min}(i)], \quad \omega_l(i) \leq \omega \leq \omega_h(i). \tag{3.33}$$

where $\hat{\Phi}_{s,\min}(i)$ is the minimum value of $\hat{\Phi}_s(\omega)$ in critical band i.

This option can be modified so that the masking threshold be explicitly used in the expression of $a(\omega)$. This second alternative is found to provide a better performance and is given by [150],

$$a(\omega) = \frac{\Phi_w(i)}{T(i)} [\Phi_w(i) + T(i)], \quad \omega_l(i) \leq \omega \leq \omega_h(i). \tag{3.34}$$

where $T(i)$ is the masking threshold in the i^{th} critical band.

To further enhance the performance, Tsoukalas *et al.* suggest to employ an iterative procedure where in every iteration a new estimate $\hat{\Phi}_s(\omega)$ of the clean speech spectrum is obtained by applying the suppression filter (3.31). This estimate serves to calculate a new masking threshold used to update $a(\omega)$. The initial estimate of $\hat{\Phi}_s(\omega)$ is still obtained using spectral subtraction. It is reported that just a couple of iterations are required to obtain a satisfactory performance of this filter, which still keeps the computational complexity of the proposed nonlinear filter at an acceptable level [150].

3.3.3 Gustafsson's method

In [56], Gustafsson *et al.* provide an alternative auditory masking based algorithm. The proposed suppression filter results in an attenuated noise which preserves the original background noise characteristics.

In this formulation, given a noisy input signal $x(n) = s(n)+w(n)$, the desired estimated signal at the output of the suppression filter can be written as

$$\hat{s}(n) = s(n) + \zeta w(n) \tag{3.35}$$

where ζ is a constant noise suppression scale factor which controls the level of the residual noise. Therefore the difference between this desired residual noise level and the actual residual noise can be quantified by the so-called residual noise distortion given by

$$E_w(\omega) = \Phi_w(\omega)(H(\omega) - \zeta)^2 \tag{3.36}$$

where $H(\omega)$ is the frequency response of the suppression filter, and $\Phi_w(\omega)$ is an estimate of the background noise PSD.

Thus forcing this residual noise distortion to be below the speech masking threshold curve, and with the constraint $0 \leq \zeta \leq H(\omega) \leq 1$, the suppression filter is found to be

$$H(\omega) = \min\left(\sqrt{\frac{T(\omega)}{\Phi_w(\omega)}} + \zeta, 1\right) \tag{3.37}$$

The masking threshold $T(\omega)$ is calculated using a mixture of Johnston's [88] and the MPEG [12] models.

3.4 The Signal Subspace Approach

The signal subspace approach (SSA) has been originally introduced as a signal processing technique for speech enhancement by Dendrinos in [32]. In his method, Dendrinos uses the singular value decomposition of a data matrix to remove the noise subspace and then reconstruct the desired speech signal from the remaining signal subspace. This approach gained more popularity when Ephraim and Van Trees proposed a new technique based on the eigenvalue decomposition of the covariance matrix of the input speech vector [41]. They

also proposed two new linear estimators based on different optimization criteria to retrieve the noise free speech from the signal subspace. In this section we analyze this technique in detail and describe the different linear signal estimators that can be used.

3.4.1 Signal and Noise Models

The speech signal can be modeled by a linear model of the form

$$\mathbf{s} = \mathbf{A}\mathbf{c} = \sum_{i=1}^{Q} \mathbf{a}_i c_i \tag{3.38}$$

where $\mathbf{s} = [s_1, \ldots, s_P]^T$ is a sequence of random signal samples and $\mathbf{c} = [c_1, \ldots, c_Q]^T$ is, in general, a zero mean complex random coefficient vector. $\mathbf{A} \in \mathbb{C}^{P \times Q}$ is a model matrix with linearly independent columns, $\mathbf{a}_i$. Therefore $\text{rank}(A) = Q \leq P$ in general. An example of such a model used with speech signals is the *damped complex sinusoid* model whose basis vector is given by [15]

$$\mathbf{a}_i = [1, \rho_i^1 e^{j\omega_i 1}, \ldots, \rho_i^{P-1} e^{j\omega_i (P-1)}]^T \tag{3.39}$$

In this thesis, the precise underlying model is not important. What is important, however, is that $Q < P$ which is a valid assumption for speech signals [41]. Hence the columns of $\mathbf{A}$ do not span the entire Euclidean space but rather a subspace referred to as the *signal subspace*. Indeed, the span of matrix $\mathbf{A}$ would be $\mathcal{R}\{\mathbf{A}\}$ as discussed in Section 3.1.1.

The covariance matrix of the vector $\mathbf{s}$ in (3.38) is given by[3]

$$\mathbf{R}_s = E\{\mathbf{s}\mathbf{s}^T\} = \mathbf{A}\mathbf{R}_c\mathbf{A}^T \tag{3.40}$$

where

$$\mathbf{R}_c = E\{\mathbf{c}\mathbf{c}^T\} \tag{3.41}$$

is the covariance matrix of vector $\mathbf{c}$, where we assume that $\mathbf{R}_c > 0$. Accordingly, $\mathbf{R}_s$ is rank deficient with $\text{rank}(\mathbf{R}_s) = Q < P$ and hence it has $P - Q$ zero eigenvalues.

Suppose now that we have available a P-dimensional noisy observation vector $\mathbf{x}$ such

[3]Unless otherwise mentioned, all signals in this thesis are considered to be real.

that

$$\mathbf{x} = \mathbf{s} + \mathbf{w} \tag{3.42}$$

where $\mathbf{w}$ is the noise vector. The noise is assumed to be zero mean, additive and uncorrelated with the speech signal. The noise covariance matrix $\mathbf{R}_w$ is assumed to be known and is given by

$$\mathbf{R}_w = E\{\mathbf{w}\mathbf{w}^T\} = \sigma^2\mathbf{I}, \tag{3.43}$$

that is the noise is a white process with variance σ^2. The whiteness assumption is necessary for the time being to be able to analyze the signal subspace method. The more practical case of colored noise will need further processing and will be addressed in Section 3.4.5. With these assumptions, the noisy signal covariance matrix $\mathbf{R}_x$ can be written as

$$\mathbf{R}_x = \mathbf{R}_s + \mathbf{R}_w = \mathbf{R}_s + \sigma^2\mathbf{I} \tag{3.44}$$

Now let $\mathbf{R}_x = \mathbf{U}\mathbf{\Lambda}_x\mathbf{U}^T$ be the eigenvalue decomposition (EVD) of $\mathbf{R}_x$. Here, the eigenvalue matrix is given by $\mathbf{\Lambda}_x = \text{diag}(\lambda_{x,1}, \ldots, \lambda_{x,P})$ with $\lambda_{x,1} \geq \lambda_{x,2} \geq \ldots \geq \lambda_{x,P}$, and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_P]$ is the matrix of orthonormal eigenvectors (i.e. $\mathbf{U}^T\mathbf{U} = \mathbf{I}$). Since the noise is white, the eigenvectors $\mathbf{u}_i$ are also the eigenvectors of $\mathbf{R}_s$ and the eigenvalues $\lambda_{x,i}$ are given by

$$\lambda_{x,i} = \begin{cases} \lambda_{s,i} + \sigma^2 & \text{for } i = 1, \ldots, Q \\ \sigma^2 & \text{for } i = Q+1, \ldots, P \end{cases} \tag{3.45}$$

where $\lambda_{s,i}$, for $i = 1, \ldots, Q$, are the Q eigenvalues of $\mathbf{R}_s$ which are strictly greater than zero.

Accordingly $\mathbf{U}$ can be partitioned as $\mathbf{U} = [\mathbf{U}_1 \quad \mathbf{U}_2]$ where $\mathbf{U}_1 = [\mathbf{u}_1, \ldots, \mathbf{u}_Q]$ and $\mathbf{U}_2 = [\mathbf{u}_{Q+1}, \ldots, \mathbf{u}_P]$. Since $\mathbf{U}$ is orthogonal we have

$$\mathbf{U}_1\mathbf{U}_1^T + \mathbf{U}_2\mathbf{U}_2^T = \mathbf{I} \tag{3.46}$$

Indeed, $\mathbf{U}_1\mathbf{U}_1^T$ is the orthogonal projector onto the subspace spanned by the columns of $\mathbf{U}_1$ which is the same as $\mathcal{R}\{A\}$. This subspace is called the *signal subspace*. $\mathbf{U}_2\mathbf{U}_2^T$, on the other hand, is the orthogonal projector onto the complementary orthogonal subspace called the *noise subspace*. It should be noted however that the noise actually fills the entire space and is not just confined to the noise subspace.

3.4.2 Linear Signal Estimation

With the signal and noise assumptions described above, a linear filter $\mathbf{H}$ is designed to estimate the desired speech vector $\mathbf{s}$ from the noisy observation $\mathbf{x}$ in (3.42). Let $\hat{\mathbf{s}}$ denote the estimate of $\mathbf{s}$ at the filter output,

$$\hat{\mathbf{s}} = \mathbf{H}\mathbf{x} = \mathbf{H}\mathbf{s} + \mathbf{H}\mathbf{w} \tag{3.47}$$

The linear estimator $\mathbf{H}$ can be calculated in different ways depending on the optimization criteria employed. We next present the most popular estimators proposed in the literature.

Least Squares Estimator (LS)

A straightforward and simple solution to the estimation problem is to use the *Least Squares* (LS) estimate. It is obtained by minimizing the squared fitting error between the observation vector $\mathbf{x}$ and the linear low order speech model of (3.38)

$$\hat{\mathbf{s}} = \mathbf{A}\mathbf{c}_0, \qquad \mathbf{c}_0 = \arg\min_{\mathbf{c}} ||\mathbf{x} - \mathbf{A}\mathbf{c}||^2 \tag{3.48}$$

Setting the gradient of the above cost function to zero, the LS solution is obtained as

$$\hat{\mathbf{s}} = \mathbf{H}\mathbf{x} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x} \tag{3.49}$$

It can be seen that $\hat{\mathbf{s}}$ is the projection of the observation vector onto the signal subspace spanned by the columns of $\mathbf{A}$ as discussed earlier. Hence $\mathbf{H}$ can alternatively be written in terms of the eigendecomposition of $\mathbf{R}_s$ as follows

$$\mathbf{H} = \mathbf{U}_1\mathbf{U}_1^T \tag{3.50}$$

This estimator does not result in any signal distortion (provided that the subspace dimension Q was correctly estimated) but has the highest possible residual noise [41]. The SNR gain obtained with this estimator is in the order of P/Q.

Other LS estimators rely on approximating the speech model matrix $\mathbf{A}$ (e.g. [83, 128, 129]), which is usually a difficult problem. Unlike these methods, (3.50) shows that such a model is not required and the desired signal can be simply estimated using the eigende-

composition of the noisy signal vector covariance matrix.

The Linear Minimum Mean Squared Error Estimator (LMMSE)

The LMMSE estimator is obtained by minimizing the residual error energy as follows

$$\min_{\mathbf{H}} E\{||\mathbf{r}||^2\} \tag{3.51}$$

where the residual error signal is defined as

$$\mathbf{r} = \hat{\mathbf{s}} - \mathbf{s} = \mathbf{H}\mathbf{x} - \mathbf{s} \tag{3.52}$$

The solution to this classical problem is given by the Wiener filter

$$\mathbf{H} = \mathbf{R}_s(\mathbf{R}_s + \sigma^2\mathbf{I})^{-1} \tag{3.53}$$

Rewriting (3.53) in terms of the EVD of $\mathbf{R}_s$ we get

$$\mathbf{H} = \mathbf{U}\mathbf{\Lambda}_s(\mathbf{\Lambda}_s + \sigma^2\mathbf{I})^{-1}\mathbf{U}^T = \mathbf{U}_1\mathbf{G}\mathbf{U}_1^T \tag{3.54}$$

where $\mathbf{G}$ is a $Q \times Q$ diagonal gain matrix with entries[4]

$$g_i = \frac{\lambda_{s,i}}{\lambda_{s,i} + \sigma^2} \quad \text{for} \quad i = 1, \ldots, Q \tag{3.55}$$

The matrix $\mathbf{U}_1^T$ is in fact the Karhunen-Loeve Transform[5] (KLT) and its effect on the noisy signal vector $\mathbf{x}$ is to calculate the coefficients of its projection onto the signal subspace. These coefficients have the property of being uncorrelated so that they can be processed independently using a diagonal gain matrix according to (3.54). The enhanced signal vector is finally reconstructed in the signal subspace using the matrix $\mathbf{U}_1$, the inverse KLT.

[4]Note the similarity between the filter in (3.55) and the Wiener filter in (3.24).

[5]To be precise, the KLT is the matrix $\mathbf{U}^T$. However since all eigenvectors in $\mathbf{U}_2^T$ will have, according to (3.55), a weight of zero, $\mathbf{U}_1^T$ can indeed be considered to be the KLT.

The Time Domain Constrained Estimator (TDC)

Instead of minimizing the total residual error energy, the Time Domain Constrained Estimator (TDC) is obtained by minimizing the signal distortion subject to forcing the residual noise energy to be below some predefined threshold. This can be achieved by decomposing the residual error signal as follows

$$\mathbf{r} = \hat{\mathbf{s}} - \mathbf{s} = (\mathbf{H} - \mathbf{I})\mathbf{s} + \mathbf{H}\mathbf{w} \tag{3.56}$$

Accordingly, define the signal distortion as

$$\mathbf{r}_s \triangleq (\mathbf{H} - \mathbf{I})\mathbf{s} \tag{3.57}$$

and

$$\mathbf{r}_w \triangleq \mathbf{H}\mathbf{w}, \tag{3.58}$$

as residual noise. The filter $\mathbf{H}$ is then obtained as the solution to the following optimization problem

$$\min_{\mathbf{H}} E\{||\mathbf{r}_s||^2\} \quad \text{subject to} \quad \frac{1}{P}E\{||\mathbf{r}_w||^2\} \leq \alpha\sigma^2 \tag{3.59}$$

where $0 \leq \alpha \leq 1$. Using the Kuhn-Tucker necessary conditions for the above constrained minimization problem [113], the optimum filter $\mathbf{H}$ is a feasible stationary point if the gradient of the Lagrangian,

$$L(H, \mu) = E\{||\mathbf{r}_s||^2\} + \mu(E\{||\mathbf{r}_w||^2\} - \alpha P\sigma^2) \tag{3.60}$$

is equal to zero and

$$\mu(E\{||\mathbf{r}_w||^2\} - \alpha P\sigma^2) = 0 \quad \text{for } \mu \geq 0. \tag{3.61}$$

The solution is then given by [41]

$$\mathbf{H} = \mathbf{R}_s(\mathbf{R}_s + \mu\sigma^2\mathbf{I})^{-1} \tag{3.62}$$

where μ is the Lagrange multiplier. The latter can be shown to satisfy the following relationship with α [41]

$$\alpha = \frac{1}{P}\text{tr}\{\mathbf{R}_s^2(\mathbf{R}_s + \mu\sigma^2\mathbf{I})^{-2}\} \tag{3.63}$$

In terms of the EVD of $\mathbf{R}_s$, the filter $\mathbf{H}$ (3.62) can be written as

$$\mathbf{H} = \mathbf{U}_1 \mathbf{G} \mathbf{U}_1^T \tag{3.64}$$

where $\mathbf{G}$ is a $Q \times Q$ diagonal gain matrix with entries

$$g_i = \frac{\lambda_{s,i}}{\lambda_{s,i} + \mu\sigma^2} \quad \text{for} \quad i = 1, \ldots, Q \tag{3.65}$$

Note that (3.65) only differs from (3.55) by the Lagrange multiplier μ, and that both are indeed the same when $\mu = 1$. Equation (3.65) can then be interpreted as a Wiener filter with a variable noise level (controlled by μ).

Equation (3.63) can also be simplified and we can find that the Lagrange multiplier satisfies

$$\alpha = \frac{1}{P} \sum_{i=1}^{Q} \left(\frac{\lambda_{s,i}}{\lambda_{s,i} + \mu\sigma^2} \right)^2 \tag{3.66}$$

The Spectral Domain Constrained Estimator (SDC)

The second estimator proposed in [41], is the spectral domain constrained approach (SDC), where the enhancement filter $\mathbf{H}$ is the solution to the following optimization problem

$$\min_{\mathbf{H}} E\{||\mathbf{r}_s||^2\}, \quad \text{subject to} \quad \begin{cases} E\{|\mathbf{u}_i^T \mathbf{r}_w|^2\} \leq \alpha_i \sigma^2 & \text{for} \quad 1 \leq i \leq Q \\ E\{|\mathbf{u}_i^T \mathbf{r}_w|^2\} = 0 & \text{for} \quad Q < i \leq P \end{cases} \tag{3.67}$$

The goal here is to minimize the signal distortion subject to keeping every spectral component of the residual noise, within the signal subspace, below some predefined threshold. Those spectral components in the noise subspace, on the other hand, are set to zero. Again using the Khun-Tucker necessary conditions, the solution to this problem is given by [41]

$$\mathbf{H} = \mathbf{U}_1 \mathbf{G} \mathbf{U}_1^T \tag{3.68}$$

where the entries of the gain matrix $\mathbf{G} = \text{diag}(g_1, \ldots, g_Q)$ are given by

$$g_i = \sqrt{\alpha_i} \quad \text{for} \quad i = 1, \ldots, Q \tag{3.69}$$

3.4.3 About the gain function

Table 3.1 The gain functions corresponding to different linear signal estimators

Signal Estimator	Gain function g_i
LS	1
LMMSE	$\frac{\lambda_{s_i}}{\lambda_{s_i}+\sigma^2}$
TDC	$\frac{\lambda_{s_i}}{\lambda_{s_i}+\mu\sigma^2}$
SDC	$\sqrt{\alpha_i}$

In theory, the gain matrix entries in (3.69) can be independent of the input data. However, exploiting information available from the signal and noise statistics may lead to a better choice of the gain coefficients. To this end, a commonly used quantity is the SNR of the i^{th} spectral component, defined as

$$\gamma_i = \lambda_{s,i}/\sigma^2 \tag{3.70}$$

Ideally, one would like to turn off spectral components with very low SNR and keep those components with very high SNR unchanged. This may be achieved by letting $g_i = f(\gamma_i)$, where $f(.)$ is an increasing function satisfying

$$\begin{aligned} f(0) &\rightarrow 0, \quad \text{and} \\ f(\infty) &\rightarrow 1 \end{aligned} \tag{3.71}$$

A possible choice of f is

$$f(\gamma) = \frac{\gamma}{\gamma + \mu} \tag{3.72}$$

leading to the TDC solution given in (3.65) (the Wiener gain function with variable noise level). A second choice is the exponential function

$$f(\gamma) = \exp(-\nu/\gamma) \tag{3.73}$$

which gives

$$g_i = \sqrt{\alpha_i} = e^{-\nu\sigma^2/\lambda_{s,i}} \quad i = 1, \ldots, Q. \tag{3.74}$$

This gain function is the one used in this thesis. This choice is motivated by the fact that this decaying exponential gain function is found to have more noise suppression capabilities. Besides, for $\nu = 1$, the first order Taylor approximation of g_i^{-1} in (3.74) is the inverse of the Wiener gain function in (3.65) with $\mu = 1$ [41].

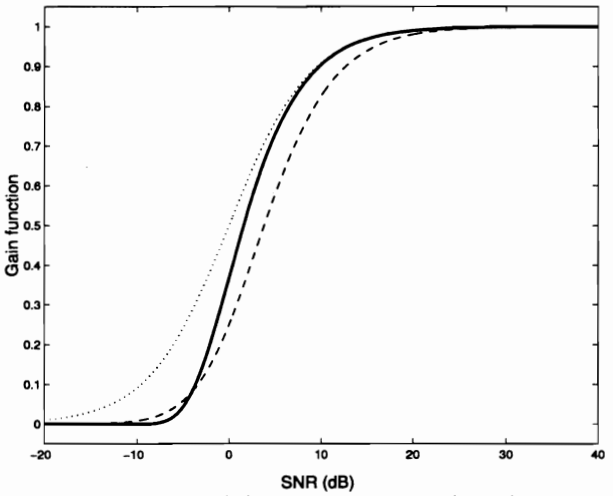


Fig. 3.2 The gain function $f(\gamma)$: exponential (3.73) with $\nu = 1$ (thick), Wiener (3.72) with $\mu = 1$ (dotted) and with $\mu = 2$ (dashed).

Figure 3.2 shows a plot of these gain functions for comparison. Note that the Least Squares (LS) estimator discussed in Section 3.4.2 is also a special case of the SDC with $g_i = 1$ for all i. The gain functions associated with the four different estimators presented are summarized in Table 3.1.

3.4.4 The SSA implementation

To implement the SSA, length-P speech vectors are input with a shift of $P/2$ samples. To preserve the whiteness of the noise, only a rectangular window is used in the analysis phase. Each of these vectors, $\mathbf{x}_n = [x(n), \ldots, x(n-P+1)]^T$, is multiplied by an enhancing linear filter $\mathbf{H}$. To synthesize the signal, the 50% overlapping enhanced vectors are then Hanning windowed and combined using the overlap-add approach [31].

Since the speech signal is not stationary over the whole utterance, the filter $\mathbf{H}$ should be updated as a new vector comes in. To this end, an estimate of the noisy speech covariance

matrix $\mathbf{R}_x$ is obtained and its EVD is calculated[6]. Using (3.45), the eigenvalues of the clean covariance matrix are estimated as follows

$$\lambda_{s,i} = \max\{\lambda_{x,i} - \sigma^2, 0\} \tag{3.75}$$

where $\lambda_{x,i}$ is the i^{th} eigenvalue of $\mathbf{R}_x$ and σ^2 is the noise variance estimated during non-speech activity periods.

In what follows we provide details about some implementation and parameter selection issues.

Estimating the covariance matrix

The linear signal estimators described earlier assume exact knowledge of the second order statistics of the noisy signal and noise process. In practice however this information needs to be estimated from the available noisy observation vectors, $\mathbf{x}_n = [x(n), \ldots, x(n-P+1)]^T$.

An estimate $\mathbf{R}_{x,n}$ of the covariance matrix of $\mathbf{x}_n$ can be obtained from the empirical covariance of $2N+1$ non-overlapping noisy vectors in the neighborhood of $\mathbf{x}_n$. To this end, we assume that conditions of stationarity and ergodicity are satisfied for a data window of length $(2N+1)P$. For speech, these conditions are considered to be satisfied for a window which is around 30 msec long [31]. The estimate $\mathbf{R}_{x,n}$ can then be obtain as follows

$$\mathbf{R}_{x,n} = \frac{1}{2PN} \sum_{i=-NP+1}^{i=NP} \mathbf{x}_{n+i}\mathbf{x}_{n+i}^T \tag{3.76}$$

$$= \mathbf{X}_n\mathbf{X}_n^T \tag{3.77}$$

where $\mathbf{X}_n$ is a $P \times 2PN$ data matrix given by

$$\mathbf{X}_n = \frac{1}{\sqrt{2PN}}[\mathbf{x}_{n-NP+1}, \ldots, \mathbf{x}_{n-1}, \mathbf{x}_n, \mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+NP}] \tag{3.78}$$

The signal subspace can now be calculated either by EVD of the covariance estimate $\mathbf{R}_{x,n}$ or via the SVD of the data matrix $\mathbf{X}_n$. Since it does not require the explicit computation of the covariance matrix, the SVD needs less computations in addition of being more stable in

[6]For simplicity of notation, we avoid the use of a hat to denote estimated quantities. Such notation will be used when it is necessary to avoid ambiguity.

the case of an ill-conditioned data matrix [50]. However, the SVD does not allow the use of more structured covariance matrices. Namely, it was observed that a Toeplitz covariance matrix would better represent speech signals and would yield a better noise reduction performance [41].

To derive such a Toeplitz covariance matrix, the biased autocorrelation function estimator obtained from $L = 2NP$ observation samples is calculated as follows

$$r_x(p) = \frac{1}{L} \sum_{i=-NP+1}^{NP-p} x(n+i)x(n+i+p) \qquad \text{for} \quad p = 0, \ldots, P-1 \tag{3.79}$$

The Toeplitz covariance matrix is then formed as follows

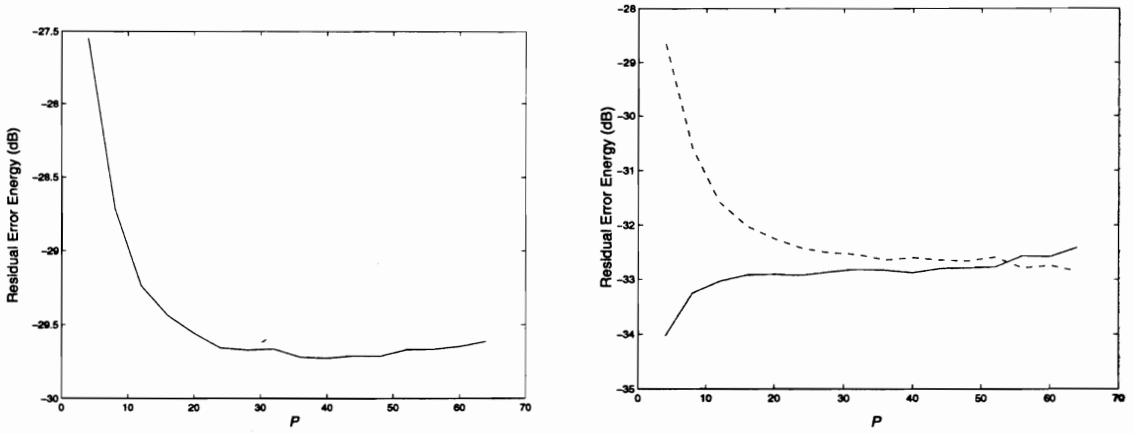
$$\mathbf{R}_x = \begin{bmatrix} r_x(0) & r_x(1) & \cdots & r_x(p-1) \\ r_x(1) & r_x(0) & \cdots & r_x(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(p-1) & r_x(p-2) & \cdots & r_x(0) \end{bmatrix} \tag{3.80}$$

The EVD of this matrix is calculated and is used to compute the signal subspace filter as described earlier.

The effect of the window length L

The choice of the window length $L = 2NP$ is a crucial design decision. To obtain better covariance estimates, L should be as long as possible. However, in the current application, we are limited by the non-stationarity of the speech signal. For this reason, we choose $L = 256$ (that is 32 msec at 8 KHz sampling rate).

Our simulations showed that for shorter windows (or frames), the covariance estimates are not reliable resulting in a higher level of the musical noise. Longer frames, on the other hand, considerably reduce the level of the residual noise at the price of more signal distortion (due to the violation of the stationarity assumption). Such distortion will be more evident at unvoiced instances of speech because they are generally shorter in duration and weaker in energy.



(a) (b)

Fig. 3.3 (a) The residual error signal and (b) the signal distortion energy (dashed) and residual noise energy (continuous), as a function of the model order P.

The effect of the model order P

Another important parameter is the model order P. Figure 3.3 (a) shows the effect of P on the total residual error energy $E\{\|\mathbf{r}\|^2\}$ while in Figure 3.3 (b) the residual noise $E\{\|\mathbf{r}_w\|^2\}$ and the signal distortion $E\{\|\mathbf{r}_s\|^2\}$ are shown separately. At low values of P, the SSA exhibits high signal distortion due to the fact that not enough correlation coefficients are available to accurately estimate the signal subspace. This results in the loss of signal components important for intelligibility. The residual noise, however, is low because, for the same reason, many of its components would have been forced to zero. The figure also shows that the higher P is, the lower the residual error energy. The latter attains a minimum value for $P > 30$ suggesting that no more gain in performance could be achieved by further increasing the value of P. Moreover, higher values of P may even increase the residual error signal energy (as can be seen in Figure 3.3 (a)) because not enough samples are available for estimating the covariance matrix (these results were obtained for a fixed frame length $L = 256$).

Besides, increasing P would drastically increase the computational load. This is because the SSA is based on the exact EVD of a $P \times P$ covariance matrix which requires $\mathcal{O}(P^3)$ floating point operations (FLOPS). Figure 3.4 shows the number of Matlab FLOPS per

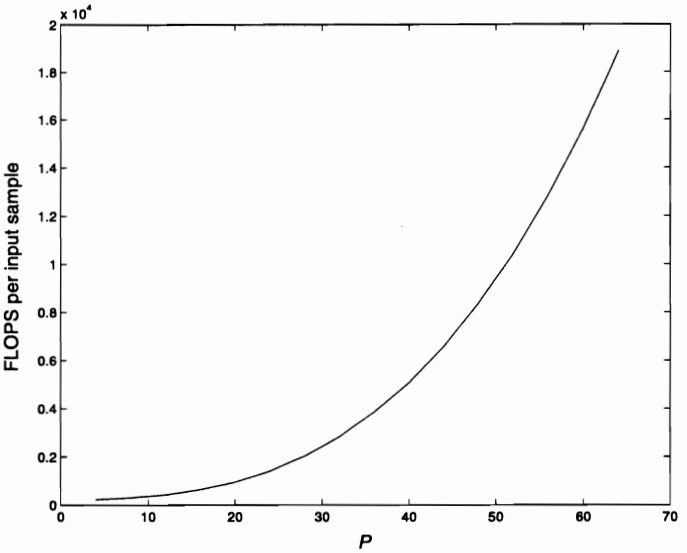


Fig. 3.4 The number of Matlab FLOPS per input sample as of function of the model order P.

input sample required by the SSA. It can be seen that the computational load increases with P.

The effect of the control parameter ν

In the exponential gain function, the parameter ν serves as a free parameter that controls the trade off between the residual noise level and the signal distortion, defined in (3.56). Figure 3.5 shows a plot of the signal distortion energy $E\{\|\mathbf{r}_s\|^2\}$, the residual noise energy $E\{\|\mathbf{r}_w\|^2\}$, and the total residual error energy $E\{\|\mathbf{r}\|^2\}$ as a function of the parameter ν in the exponential gain function (3.74). It can be seen that as ν increases, the signal distortion increases and the residual noise level decreases. Consequently, the minimum values for the total residual error energy is obtained when ν is around 1.5. Listening tests however show that $\nu = 2$ is a better choice from a perceptual perspective. This can be explained by the fact that humans prefer a lower noise level at the expense of more signal distortion. Note that since the noise and the speech signal are uncorrelated, $E\{\|\mathbf{r}\|^2\} = E\{\|\mathbf{r}_s\|^2\} + E\{\|\mathbf{r}_w\|^2\}$.

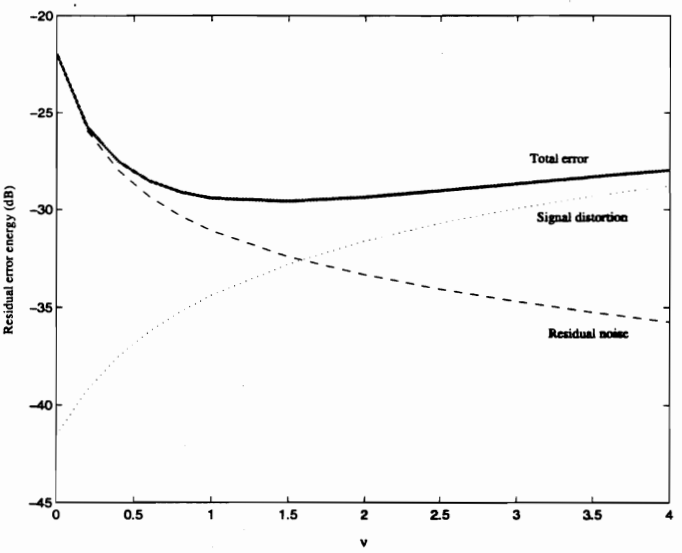


Fig. 3.5 The total residual error signal energy (thick), the signal distortion energy (dotted) and the residual noise energy (dashed), as a function of ν.

3.4.5 Handling colored noise

One problem with the signal subspace approach is that it is based on the white noise assumption. However, almost all common noise types encountered in real life are colored. Therefore extra techniques should be included with the signal subspace method to handle the colored noise case for it to be useful in practice. Fortunately, several such techniques have been proposed in the literature with satisfying results.

Prewhitening

In [41], prewhitening is proposed as a remedy to the colored noise case. It consists of multiplying the noisy input vector $\mathbf{x}$ by $\mathbf{R}_w^{-\frac{1}{2}}$, the square root of the colored noise covariance matrix $\mathbf{R}_w = E\{\mathbf{w}\mathbf{w}^T\}$. The prewhitened signal is obtained as

$$\breve{\mathbf{x}} = \mathbf{R}_w^{-\frac{1}{2}}\mathbf{x} = \mathbf{R}_w^{-\frac{1}{2}}\mathbf{s} + \mathbf{R}_w^{-\frac{1}{2}}\mathbf{w} = \breve{\mathbf{s}} + \breve{\mathbf{w}} \tag{3.81}$$

It can be verified that $E\{\breve{\mathbf{w}}\breve{\mathbf{w}}^T\} = \mathbf{I}$. Hence $\breve{\mathbf{w}}$, the prewhitened noise component, is now white with variance equal to one. This can be seen in Figure 3.6 where the spectrum of a Volvo car noise signal sampled at 8 KHz, has been whitened yielding the relatively flat unit variance (0 dB) spectrum shown in the figure. The power spectra were obtained using

the Blackman-Tukey estimator with a Bartlett window. In Figure 3.7, the same result is shown for an F16 jet cockpit noise. It can also be seen that the noise spectrum has been flattened but not to the same degree as in the Volvo noise case.

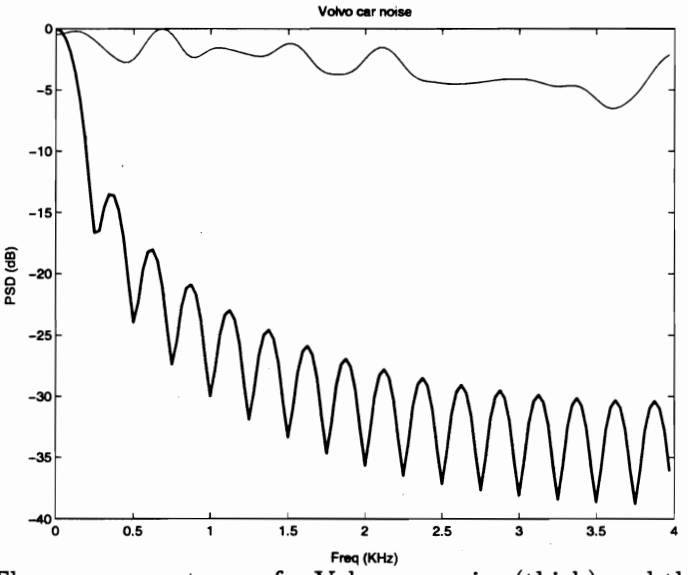


Fig. 3.6 The power spectrum of a Volvo car noise (thick) and the spectrum of the corresponding prewhitened signal (thin).

The EVD obtained from the signal $\breve{\mathbf{x}}$ can now be used instead of the EVD obtained from $\mathbf{x}$ to calculate a filter $\breve{\mathbf{H}}$ using any of the linear estimators presented earlier. However, since the desired speech signal is also affected, the inverse of the prewhitening matrix, $\mathbf{R}_w^{\frac{1}{2}}$, is applied as a postfilter to undo the effect of prewhitening. This is called dewhitening. Accordingly, the overall effective enhancing filter becomes

$$\overline{\mathbf{H}} = \mathbf{R}_w^{\frac{1}{2}} \breve{\mathbf{H}} \mathbf{R}_w^{-\frac{1}{2}} \tag{3.82}$$

The prewhitening and dewhitening matrices can be obtained using the Cholesky decomposition of the noise covariance matrix or more safely (in case the latter is not invertible or near singular) using its eigenvalue decomposition. Consider the EVD $\mathbf{R}_w = \mathbf{U}_w \mathbf{\Lambda}_w \mathbf{U}_w^T = \mathbf{U}_{w,1} \mathbf{\Lambda}_{w,1} \mathbf{U}_{w,1}^T$, where $\mathbf{\Lambda}_{w,1}$ contains only non-zero eigenvalues and $\mathbf{U}_{w,1}^T$ has the correspond-

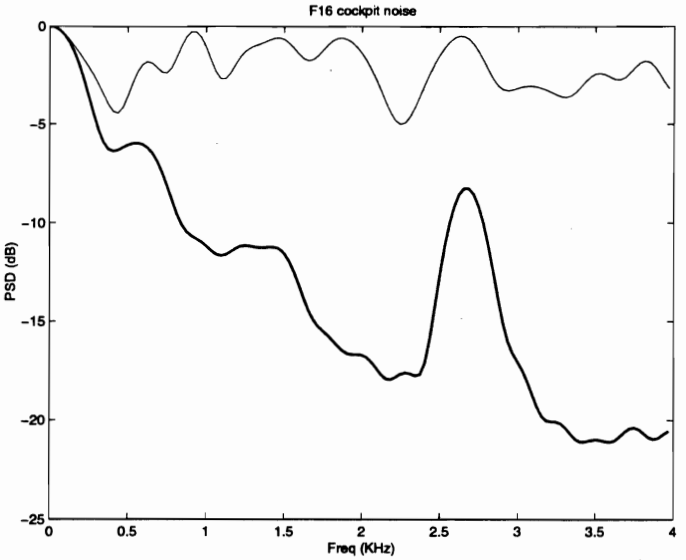


Fig. 3.7 The power spectrum of an F16 jet cockpit noise (thick) and the spectrum of the corresponding prewhitened signal (thin).

ing eigenvectors as its columns, then

$$\mathbf{R}_w^{\frac{1}{2}} = \mathbf{U}_{w,1}\mathbf{\Lambda}_{w,1}^{\frac{1}{2}}\mathbf{U}_{w,1}^T \tag{3.83}$$

$$\mathbf{R}_w^{-\frac{1}{2}} = \mathbf{U}_{w,1}\mathbf{\Lambda}_{w,1}^{-\frac{1}{2}}\mathbf{U}_{w,1}^T \tag{3.84}$$

We shall refer to this method as the PreWhitening based Signal Subspace method (PWSS). The effect of prewhitening will be further addressed and analyzed in Section 4.3.

In [69, 70] prewhitening is accomplished using a filter designed from the coefficients of an autoregressive model of the noise. Suppose that the noise signal can be modeled by an AR process of order q,

$$w(n) = -\sum_{i=1}^{q} a(i)w(n-i) + v(n) \tag{3.85}$$

where $v(n)$ is a white Gaussian process with variance σ^2. Then after estimating the AR parameters, using the modified covariance method [94], the prewhitening filter impulse response is obtained as follows,

$$h(n) = \begin{cases} \hat{a}(n) & \text{if } 0 \leq n \leq q \\ 0 & \text{otherwise,} \end{cases} \tag{3.86}$$

where $\hat{a}(n)$ are the estimates of the AR model parameters and $\hat{a}(0) = 1$. The inverse filter impulse response, say $h^{-1}(n)$, on the other hand is given by

$$h^{-1}(n) = -\sum_{i=1}^{q} \hat{a}(i)h^{-1}(n-i) + \delta(n), \quad \text{for } n = 0, 1, \ldots \tag{3.87}$$

where $\delta(0) = 1$ and $\delta(n) = 0$ for $n \neq 0$.

The generalized eigenvalue decomposition method

Prewhitening can alternatively be realized as an integral part of the subspace decomposition using the generalized EVD [68] or the generalized SVD [84]. The idea is to find a matrix that would diagonalize both $\mathbf{R}_s$ and $\mathbf{R}_w$ simultaneously. Such a matrix would satisfy [68],

$$\mathbf{V}^T\mathbf{R}_s\mathbf{V} = \mathbf{\Lambda} \tag{3.88}$$

$$\mathbf{V}^T\mathbf{R}_w\mathbf{V} = \mathbf{I} \tag{3.89}$$

where $\mathbf{V}$ and $\mathbf{\Lambda}$ are the eigenvector matrix and the eigenvalue matrix of $\mathbf{R}_w^{-1}\mathbf{R}_s$, respectively. Hence the optimal filter (3.68) can be modified as follows

$$\overline{\mathbf{H}} = \mathbf{V}^{-T}\mathbf{G}\mathbf{V}^T \tag{3.90}$$

It should be noted that $\mathbf{V}^T$ is no longer the KLT corresponding to $\mathbf{R}_s$ and that $\mathbf{V}$ is not orthogonal. The gain matrix $\mathbf{G}$ is chosen as discussed earlier to satisfy the desired optimization criterion. The noise variance, however, should now be set to one, that is $\sigma^2 = 1$.

The Raleigh Quotient method

As discussed earlier, the prewhitening technique consists of using $\breve{\mathbf{x}}$, in (3.81), instead of $\mathbf{x}$ for the filter design. Therefore, the filter will shape the noise spectrum according to the spectrum of $\breve{\mathbf{s}}$, the modified speech vector, rather than $\mathbf{s}$. Hence the filter in equation (3.82) is not necessarily optimal in the sense of its noise shaping capabilities [121].

Alternatively, another method to handle colored noise, consists of replacing the constant

noise variance in (3.70) by the noise energy in the direction of the i^{th} eigenvector, given by

$$\xi_i = \mathbf{u}_i^T \mathbf{R}_w \mathbf{u}_i \tag{3.91}$$

which is the Raleigh quotient associated with $\mathbf{u}_i$ and $\mathbf{R}_w$ for $i = 1, \ldots, Q$. Here $\mathbf{u}_i$ is the i^{th} eigenvector of the clean covariance matrix estimate, $\hat{\mathbf{R}}_s$, with corresponding eigenvalue $\lambda_{s,i}$. $\hat{\mathbf{R}}_s$ is estimated from the noisy covariance matrix as follows

$$\hat{\mathbf{R}}_s = \mathbf{R}_x - \mathbf{R}_w \tag{3.92}$$

Since $\hat{\mathbf{R}}_s$, so obtained, is no longer guaranteed to be positive definite, the rank Q is chosen as the number of strictly positive eigenvalues $\lambda_{s,i}$. The gain function is calculated, for example using the exponential function (3.74), in the following way,

$$g_i = f(\lambda_{s,i}/\xi_i) = e^{-\nu \xi_i / \lambda_{s,i}} \quad \text{for } i = 1, \ldots, Q \tag{3.93}$$

This method, which we will refer to as the Raleigh Quotient Signal Subspace method (RQSS), was found to be superior to the prewhitening technique in the sense that better noise shaping is achieved [121, 130]. This method also requires less computations than PWSS because no matrix inversion is involved. RQSS was the basis for the method described in [121] and [130]. In the latter it was used in conjunction with a subspace tracking technique in order to reduce the computational load.

In [121] further processing is added to RQSS by classifying the speech frames as *speech dominated* or *noise dominated*. The procedure described above is applied during speech dominated frames. During noise dominated frames on the other hand, the EVD of the noise covariance matrix is used instead of that of the estimated clean speech covariance matrix. This alternative scheme is described next.

Let $\mathbf{R}_w = \mathbf{U}_w \mathbf{\Lambda}_w \mathbf{U}_w^T$ be the EVD of $\mathbf{R}_w$ where $\mathbf{U}_w = [\mathbf{u}_{w,1} \ \ldots \ \mathbf{u}_{w,P}]$ and $\mathbf{\Lambda}_w = \text{diag}\{\lambda_{w,1}, \ldots, \lambda_{w,P}\}$ are the corresponding eigenvector and eigenvalue matrices respectively. The gain coefficients are now given by

$$g_i = f(\phi_i/\lambda_{w,i}) = e^{-\nu \lambda_{w,i} / \phi_i} \quad \text{for } i = 1, \ldots, Q \tag{3.94}$$

where $\phi_i = \mathbf{u}_{w,i}^T \mathbf{R}_x \mathbf{u}_{w,i} - \lambda_{w,i}$, for $i = 1, \ldots, Q$. That is ϕ_i is the speech energy estimate

in the direction of the i^{th} eigenvector of $\mathbf{R}_w$ obtained by subtracting the noise energy from the noisy speech energy along that direction. The rank is again chosen such that $\phi_i > 0$ for $i = 1, \ldots, Q$. However, our experiments show that while this method indeed outperforms the PWSS method, the main gain in performance is due to the Raleigh quotient technique rather than the frame classification approach.

The RQSS will be used to evaluate the merit of using the human hearing properties in the novel Perceptual SSA method, which will be presented in Chapter 5. It will also be generalized into a multi-microphone design in Chapter 6, leading to another novelty of this thesis. RQSS basically represents the methods in [130] and [121] where our experiments revealed that the superiority of these methods over the original SSA with prewhitening can mainly be attributed to the use of the Raleigh Quotient approach.

3.5 Noise Estimation

For all single channel speech enhancement methods, noise estimation remains a very challenging problem. Indeed, all these methods rely on an accurate estimation of the background noise for an acceptable performance. Inaccurate estimation of the noise usually results in the suppression of important speech components or the increased intensity of the musical noise. For this reason, much research effort has been put to improve noise estimation methods though still non of the proposed approaches is fully satisfactory.

Most common methods require a voice activity detector (VAD) to make a hard decision on the presence or absence of speech in the input speech frame. Based on that decision the noise estimate is updated using a first order recursive system as follows

$$\Phi_w(\omega, m) = \rho\Phi_w(\omega, m-1) + (1-\rho)\Phi_x(\omega, m) \tag{3.95}$$

where $\hat{\Phi}_w(\omega, m)$ and $\Phi_x(\omega, m)$ are the noise PSD estimate and the input noisy speech PSD of frame m respectively. The parameter $0 \leq \rho \leq 1$ is a forgetting factor selected to adjust the sensitivity of the noise update scheme to new input data. For example during speech activity, ρ is set to one so that the previous noise estimate is kept unchanged. In the context of SSA, an estimate of the noise covariance matrix is updated in a similar way

$$\mathbf{R}_{w,m} = \rho\mathbf{R}_{w,m-1} + (1-\rho)\mathbf{R}_{x,m} \tag{3.96}$$

where $\hat{\mathbf{R}}_{w,m}$ and $\mathbf{R}_{x,m}$ are the noise covariance matrix estimate and the noisy speech covariance matrix for the m^{th} frame respectively.

This approach has been found useful and is widely used in practice. However its effectiveness is highly dependent on the accuracy of the VAD. Otherwise, if a speech frame is mistakenly labeled as a speech free frame, then the noise estimate becomes inaccurate resulting in the suppression of important speech components. Indeed, the assumption that speech and noise are uncorrelated, on which almost all noise reduction methods are based, would be violated.

3.5.1 Voice activity detection

Voice activity detection is unfortunately a very difficult problem especially at low SNR conditions. Over the years, several methods have been developed which mainly rely on extracting some measured features and comparing them with thresholds to decide on the presence or absence of speech. For non-stationary noise, these thresholds have to be time varying.

The most popular VAD methods are based on the energy of the input signal. This energy, in the presence of speech, is believed to be higher than the background noise. Therefore, if the calculated energy is above a predefined threshold then the current analysis frame is labeled has a speech active frame [90, 89, 165]. Other methods include those based on zero crossing [90], a periodicity measure [151], cepstral coefficients [11] and adaptive noise modeling [168]. A fusion of two (or more) of these measures has also been proposed [147].

In the adaptive signal subspace approach proposed in [130], a voice activity detector based on the principal component of noisy speech, i.e the largest eigenvalue $\lambda_{x,1}$ of the covariance matrix has been proposed. As will be seen in Section 4.2, the largest eigenvalue corresponds to the energy of the first speech formant. The method consists of tracking the minimum and maximum of the principal component and setting a threshold value to be $1/12th$ the distance between them. Voice activity is detected if $\lambda_{x,1}$ is greater than the minimum value by at least that threshold.

We note also that lately, in [142], it was reported that voice activity detection errors can be tolerated if they occur at a rate of no more than 20% of the time.

3.5.2 Quantile based noise estimation

Despite the variety of VAD techniques developed, the latter still do not offer satisfactory results especially under very noisy conditions. Non-stationary noise also poses a serious obstacle as it necessitates updating the decision thresholds on the fly, which is not a straightforward task.

Alternatively, some researchers have recently proposed to estimate the noise continually even during speech activity making the presence of a VAD unnecessary. One such approach, is the so called quantile based noise estimation method (QBNE). Originally proposed in [143], this method is actually an extension of the histogram approach of [66]. This technique is driven by the assumption that the noise is stationary or at least its statistics are changing slower than those of the clean speech. This assumption is frequently encountered in many real life situations. In fact it is known that even during speech activity, the speech signal does not permanently occupy all frequency bands. Accordingly, it is possible to assume that for a long enough period of time, the energy per frequency band is at the noise level. Hence the noise estimate $\Phi_w(\omega, m)$ at the m^{th} frame is obtained in the following way.

Let the current and previous $T-1$ noisy speech frames $\Phi_x(\omega, t)$, for $t = m-T+1, \ldots, m$, be stored in a length T buffer. These PSD's are then sorted in an ascending order such that

$$\Phi_x(\omega, t_1) \leq \Phi_x(\omega, t_2) \leq \ldots \leq \Phi_x(\omega, t_T) \tag{3.97}$$

where $t_j \in [m - T + 1, m]$. The q^{th} quantile for every frequency is taken as the noise estimate for the current frame at that frequency,

$$\Phi_w(\omega) = \Phi_x(\omega, t_{\lfloor qT \rfloor}) \tag{3.98}$$

where $\lfloor . \rfloor$ denotes flooring to the nearest integer. That is, $q = 0$ gives the minimum, $q = 1$ gives the maximum and $q = 0.5$ gives the median. In [118] the minimum has been used whereas in [42, 43, 143] it was reported that the median has a better performance as it is less vulnerable to outliers. In the case of the median, the underlying assumption is that for T frames, one particular frequency is occupied by a speech component in at most 50 % of the time. A closely related approach based on minima tracking for recursive noise estimation is also reported in [22] and [33].

3.6 Summary

In this chapter we provided a survey of some of the single channel speech enhancement methods. For the frequency domain methods we focused on Wiener filtering and spectral subtraction and we discussed their main limitation, namely, the musical noise. As a remedy to this artifact, we examined in detail three methods based on auditory masking which were reported to make the musical noise less annoying with no increased signal distortion.

We then provided a detailed presentation of the signal subspace approach for speech enhancement which is the basic method in this thesis. Unlike its frequency domain counterparts, the SSA performs in a signal dependent domain controlled by a KLT derived from the signal covariance matrix. We examined the sensitivity of the method to changes in the design parameters and we examined the modifications proposed in literature to generalize the SSA so that it can handle colored noise situations. The SSA is reported to yield improved noise reduction performance in general though at the expense of increased complexity.

Finally, we addressed the problem of noise estimation which is very crucial to all single channel speech enhancement methods. Usually, these methods rely on a VAD to continue updating a noise estimate until speech presence is detected at which point the update is frozen. We briefly discussed some VAD techniques and then presented a relatively newer alternative where noise estimation can still be carried out even during speech activity.

Chapter 4

The Frequency to Eigendomain Transformation

Like most single channel speech enhancement methods, the SSA, presented in Section 3.4, suffers from the annoying residual noise known as *musical noise*. Tones at random frequencies, created due to poor estimation of the signal and noise statistics, are at the origin of this artifact. In Section 3.3, we have presented some methods that propose to solve this problem by exploiting the human masking properties. These methods are based on the fact that the human auditory system is able to tolerate additive noise as long as it is below some *masking threshold*. It has been reported that this approach allowed to reduce the intensity of the musical noise [5, 25, 56, 150, 157].

Recently, a DCT based SSA imitating the human hearing resolution was proposed [156]. However, no algorithm which employs a sophisticated hearing model with a KLT based SSA is available[1]. The reason is that the SSA does not operate in the frequency domain where the available masking models are developed. Indeed, as discussed in Chapter 2, almost all existing masking models were developed in the frequency domain mainly because the human hearing properties were studied (hence understood) as a function of frequency. Therefore, any attempt to use auditory masking to enhance the performance of SSA, should first identify a way to map the human hearing properties from the frequency domain to the eigendomain.

In this chapter, we adopt two known relationships in signal processing and linear algebra

[1]Lately, however, after we published parts of our work in [76] and [78], some methods addressing this issue have emerged, for example [99] and [97].

that relate the EVD of a signal covariance matrix to its PSD and vice versa [53, 64]. These relationships, which we will refer to as the Frequency to Eigendomain Transformation (FET) and its "inverse" (IFET), will be used in the context of speech enhancement to map the auditory masking properties from the frequency domain to the eigen-domain. This, as will be shown in Chapter 5, allows to design a signal subspace filter which takes advantage of the masking properties to yield better residual noise shaping from a psychoacoustic perspective. Recall that noise shaping refers to modifying the residual noise spectrum in a way that it takes the shape of the desired speech signal hence making it less audibale due to masking effects.

Our new method, referred to as the Perceptual Signal Subspace (PSS) method and which we introduce in Chapter 5, uses the IFET to obtain a PSD estimate of the speech signal which is used to compute a masking threshold. This threshold is mapped to the eigendomain using the FET relationship and a modified gain function is calculated accordingly yielding the PSS method.

We also use the FET to provide an analysis of the SSA according to a filterbank interpretation. This interpretation helps to view the SSA from a different angle and to explain its effect on the input signal PSD. Accordingly, we discuss and try to explain some SSA related observations reported in the literature. This will also serve to motivate our choice of the gain function for the proposed PSS method.

4.1 Derivation

Consider a real zero mean WSS stochastic process $x(n)$ with autocorrelation function

$$\tilde{r}(p) = E\{x(n)x(n+p)\}$$

The PSD of $x(n)$ is defined as follows

$$\tilde{\Phi}(\omega) = \sum_{p=-\infty}^{\infty} \tilde{r}(p)e^{-j\omega p} \tag{4.1}$$

In practice, however, we need to estimate the PSD from a single realization of $x(n)$ over a finite time interval, say $n \in [0, \ldots, L-1]$. To this end, consider the biased autocorrelation

estimator given by

$$r(p) = \frac{1}{L} \sum_{n=0}^{L-1-p} x(n)x(n+p) \quad p = 0, \ldots, L-1 \tag{4.2}$$

with $r(-p) = r(p)$ and $r(p) = 0$ for $|p| \geq L$. The PSD can then be estimated using the periodogram defined as [64]

$$\Phi(\omega) = \sum_{p=-L+1}^{L-1} r(p) e^{-j\omega p} \tag{4.3}$$

Now let $\mathbf{R} = \text{Toeplitz}\big(r(0), \ldots, r(P-1)\big)$ be the covariance matrix estimate of $x(n)$ with λ_i being its i^{th} eigenvalue and $\mathbf{u}_i = [u_i(0), \ldots, u_i(P-1)]^T$ being the corresponding unit-norm eigenvector. The rank of $\mathbf{R}$ is in general $Q \leq P$, so that $\lambda_i = 0$ for $i > Q$.

Property 4.1.1 *The eigenvalues λ_i, for $i = 1, \ldots, Q$, can be written in terms of $\Phi(\omega)$ in the following way*

$$\lambda_i = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi(\omega) |V_i(\omega)|^2 d\omega \quad \textit{for} \quad i = 1, \ldots, Q \tag{4.4}$$

where

$$V_i(\omega) = \sum_{p=0}^{P-1} u_i(p) e^{-j\omega p} \tag{4.5}$$

is the Discrete-Time Fourier Transform of the entries $u_i(p)$ of the eigenvector $\mathbf{u}_i$.

Proof: By definition the eigenvalue λ_i can be written as

$$\begin{aligned} \lambda_i &= \mathbf{u}_i^H \mathbf{R} \mathbf{u}_i \\ &= \sum_{p=0}^{P-1} \sum_{q=0}^{P-1} u_i^*(p) r(p-q) u_i(q) \end{aligned} \tag{4.6}$$

Using the relationship between the autocorrelation function estimate and the periodogram, i.e. the inverse DTFT of (4.3):

$$r(p) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi(\omega) e^{j\omega p} d\omega \tag{4.7}$$

we have

$$\lambda_i = \sum_{p=0}^{P-1}\sum_{q=0}^{P-1} u_i^*(p)u_i(q)\frac{1}{2\pi}\int_{-\pi}^{\pi}\Phi(\omega)e^{j\omega(p-q)}d\omega \tag{4.8}$$

Recalling the definition of $V_i(\omega)$ (4.5) and rearranging terms we get

$$\lambda_i = \frac{1}{2\pi}\int_{-\pi}^{\pi}\Phi(\omega)V_i(\omega)V_i^*(\omega)d\omega \tag{4.9}$$

Which completes the proof.

■

In this thesis (4.4) is called the Frequency to Eigendomain Transformation (FET). This relationship constitutes a mapping from the frequency domain to the eigendomain. Therefore, if we have available a masking threshold $T(\omega)$ calculated using one of the masking models described in Section 2.3, then this threshold can be mapped to the eigendomain using the FET. The so obtained eigenvalues, or more accurately "masking energies", will reflect the masking properties associated with the currently processed speech frame.

Any masking model, however, requires an estimate of the clean speech PSD in order to calculate the masking threshold. To this end, assuming that the clean speech estimate is available in the eigen-domain, we need a second relationship, or a sort of inverse FET, which maps the available EVD to the frequency domain. Such relationship can be obtained using the Blackman-Tukey PSD estimator.

As discussed in Section 3.1.3, the Blackman-Tukey estimator can be obtained by multiplying $r(p)$ in (4.3) by a length $2P-1$ window $w_b(p)$, where $P \leq L$ as follows

$$\Phi_B(\omega) = \sum_{p=-P+1}^{P-1} r(p)w_b(p)e^{-j\omega p} \tag{4.10}$$

Property 4.1.2 *If $w_b(p)$ is a Bartlett (triangular) window, then $\Phi_B(\omega)$ can be written in terms of the eigenvalue decomposition of $\mathbf{R}$ in the following way [64, 94]*

$$\Phi_B(\omega) = \frac{1}{P}\sum_{i=1}^{Q}\lambda_i|V_i(\omega)|^2 \tag{4.11}$$

Proof: Consider the Blackman-Tukey estimate (4.10), assuming a triangular window

(i.e. $w_b(p) = 1 - \frac{|p|}{P}$ for $|p| < P$), we get

$$\Phi_B(\omega) = \frac{1}{P} \sum_{p=-P+1}^{P-1} r(p)(P - |p|)e^{-j\omega p} \tag{4.12}$$

The above summation over p is readily expressible as a double summation as

$$\Phi_B(\omega) = \frac{1}{P} \sum_{p=0}^{P-1} \sum_{q=0}^{P-1} r(p-q)e^{-j\omega(p-q)} \tag{4.13}$$

From the eigenvalue decomposition formula (3.2),

$$\mathbf{R} = \sum_{i=1}^{P} \lambda_i \mathbf{u}_i \mathbf{u}_i^H,$$

we note that

$$r(p-q) = \sum_{i=1}^{P} \lambda_i u_i(p) u_i^*(q) \tag{4.14}$$

Substituting (4.14) into (4.13) and recalling the definition of $V_i(\omega)$ (4.5), we finally obtain

$$\Phi_B(\omega) = \frac{1}{P} \sum_{i=1}^{Q} \lambda_i |V_i(\omega)|^2 \tag{4.15}$$

where the limit of the summation is changed from P to Q because $\lambda_i = 0$ for $i > Q$.

■

The FET and IFET relationships developed here are intended to be used as the inverse of each other. However, it should be stated that mathematically speaking, this is not true in general. In fact, while the Blackman-Tukey PSD estimate $\Phi_B(\omega)$ can be expressed in terms of the EVD of the covariance matrix of the signal $x(n)$ as in the IFET relationship (4.11), inserting it instead of $\Phi(\omega)$ in the FET relationship (4.4), will not yield the signal eigenvalues λ_i's, that is

$$\lambda_i \neq \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_B(\omega) |V_i(\omega)|^2 d\omega \tag{4.16}$$

in general[2].

[2]Equality is obtained if the signal $x(n)$ was white.

4.2 A filterbank interpretation

By design, the SSA is always viewed and analyzed from a linear algebra perspective. However since our understanding of speech signals is best in terms of its frequency spectrum, it seems beneficial if we can provide a frequency domain interpretation of the SSA in order to better understand its behaviour. A filterbank interpretation has been given for example in [60] and [85] yielding a modified SSA based method (with an SVD implementation). We further pursue this interpretation here using the FET and consequently try to explain some phenomena related to SSA.

In this section, all signals are sampled at 8 KHz hence have a bandwidth of 4 KHz. The speech power spectra estimates shown in the figures, are obtained using the Blackman-Tukey estimate calculated using a length $2P-1$ Bartlett window, with $P=32$, from a length $L=256$ frame.

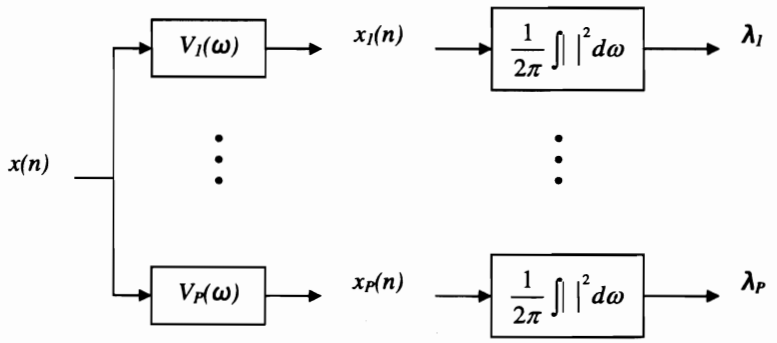


Fig. 4.1 A block diagram of the filterbank interpretation of the FET

Consider a filter bank with P analysis filters with frequency responses $V_i(\omega)$ for $i=1,\ldots,P$ as shown in Figure 4.1. That is, every filter has a finite impulse response $u_i(p)$ for $p=0,\ldots,P-1$. Now let $x(n)$, a random process with PSD $\Phi(\omega)$, be the input to this filterbank. Thus, the PSD $\Phi_i(\omega)$ of the output $x_i(n)$ at the i^{th} filter is given by [65]

$$\Phi_i(\omega)=\Phi(\omega)|V_i(\omega)|^2 \tag{4.17}$$

Using FET (4.4), it can be seen that the total energy at the output of the i^{th} filter is

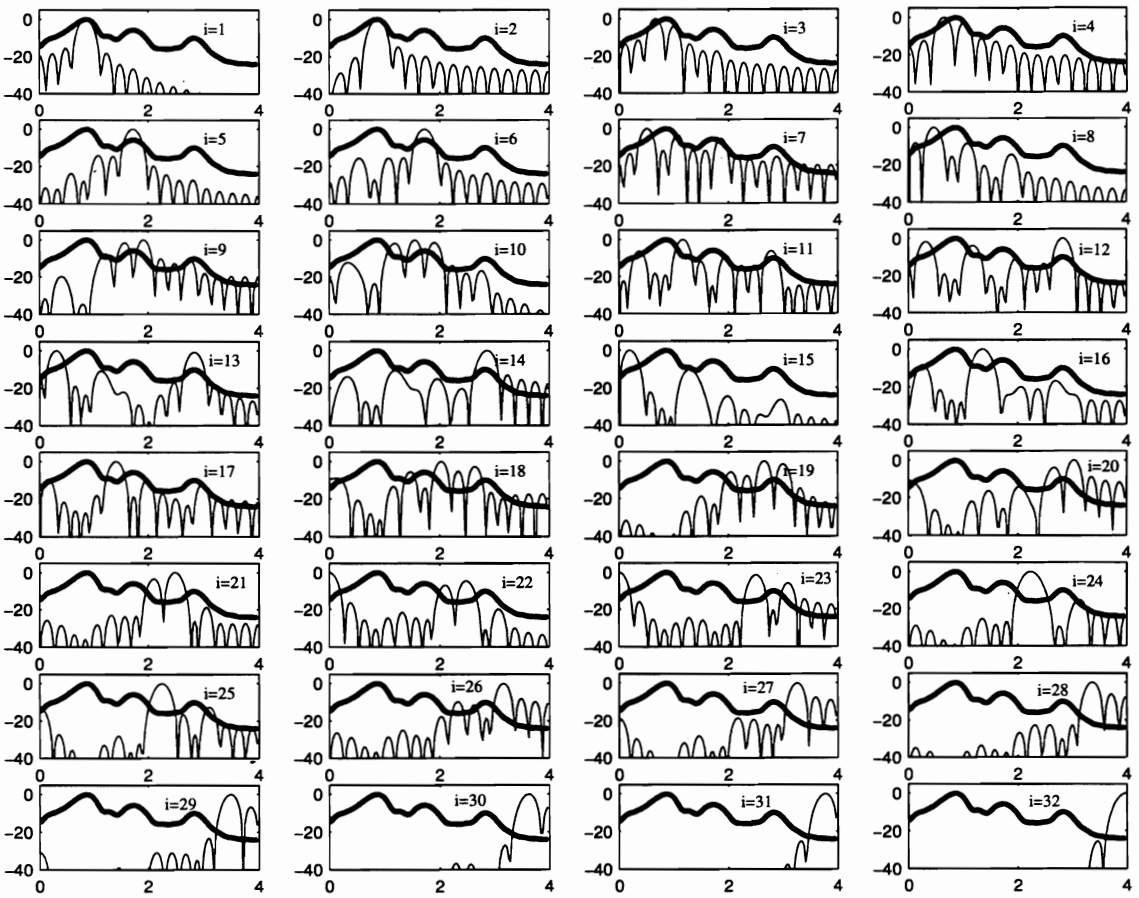


Fig. 4.2 The magnitude squared of the i^{th} eigenfilter $V_i(\omega)$ for the vowel /a/ in the word "cats", for $i = 1, \ldots, P$. The thick line shows the PSD of the speech signal.

actually the i^{th} eigenvalue λ_i,

$$\lambda_i = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_i(\omega) d\omega \tag{4.18}$$

Therefore, the SSA actually consists of dividing the input signal into several frequency bands. In every band, a gain function depending on the average SNR in that particular band is applied and then the whole signal is re-synthesized in the time domain.

This filterbank, however, is different from other common ones, such as the DFT filter banks, in that instead of having the passbands of the analysis filters uniformly distributed over the frequency range of interest, the "eigen" analysis filters are signal dependent. In Figures 4.2 and 4.3, these filters are shown for the case of a vowel (/a/) and an affricate (/ch/), respectively. The figures show the PSD of the input signal (thick line) together

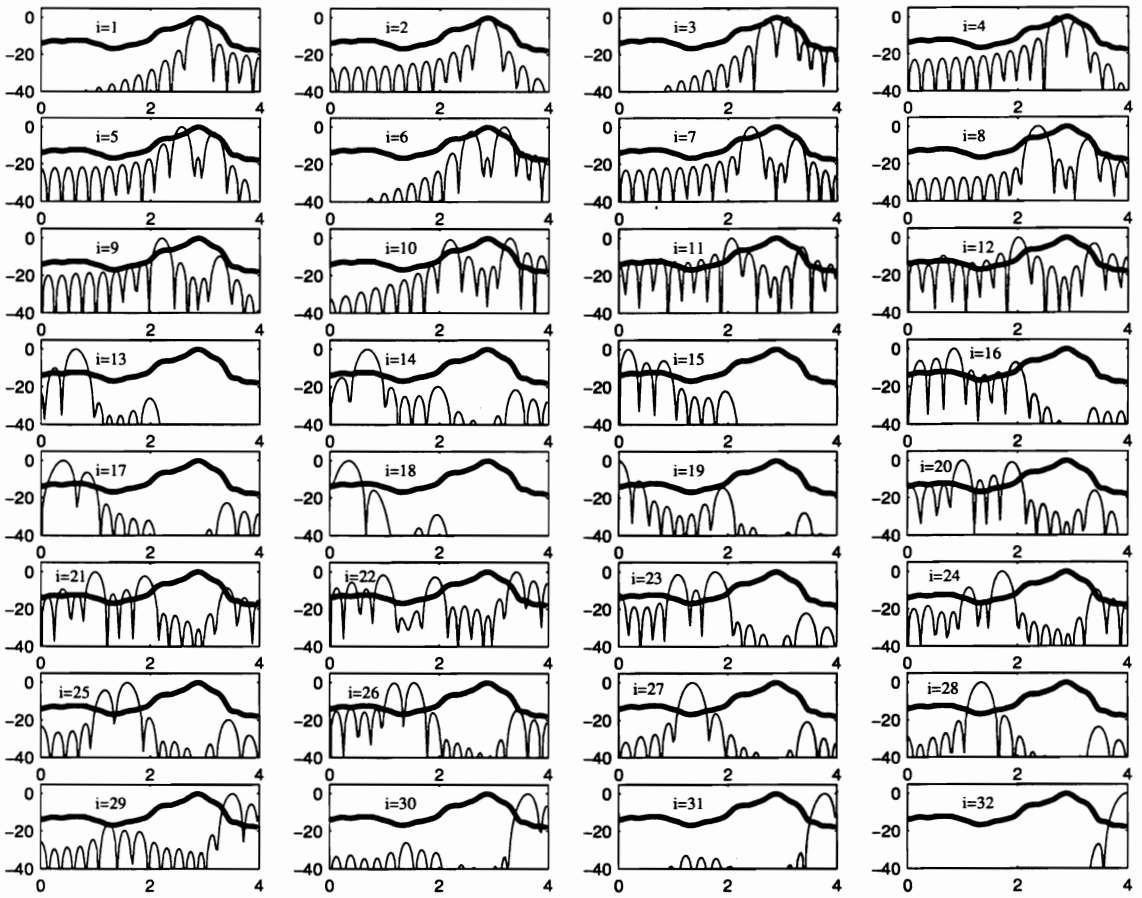


Fig. 4.3 The magnitude squared of the i^{th} eigenfilter $V_i(\omega)$ for the affricate /ch/ in the word "each", for $i = 1, \ldots, P$. The thick line shows the PSD of the speech signal.

with the magnitude squared of the frequency response of the $P = 32$ eigen analysis filters, $|V_i(\omega)|^2$.

In Figure 4.2, for the vowel, it can be seen that the first four filters correspond to the first formant whereas the next two filters correspond to the second formant. The third formant (also important for intelligibility) can be found in the pass-bands of the 12^{th}, 13^{th} and 14^{th} filters. The passbands of the analysis filters corresponding to the affricate /ch/ are also shown in Figure 4.3. It can also be seen that the passbands of the filters corresponding to the largest eigenvalues coincide with the first formant of this speech signal.

As mentioned in Section 3.2, the output of the single channel frequency domain methods usually suffer from spectral peaks randomly distributed over time and frequency which are

commonly referred to as musical noise. This artifact mostly occurs due to poor estimation of the speech and noise statistics resulting in "random" fluctuations in the suppression filters both over time and frequency. Therefore, the proposed remedy to this problem usually consists of trying to smooth the filter coefficients. The use of masking in speech enhancement can be also viewed as smoothing in the frequency domain by applying some perceptual criteria. In the SSA, and using the filterbank interpretation, it can be readily noted that this approach accomplishes such smoothing by obtaining the average SNR within every passband of the eigen analysis filters. The reduction of the musical noise, commonly reported for SSA based methods, can actually be attributed to this phenomenon.

In addition to that, and since the passbands of the analysis filters are usually located around the speech formants, the residual noise spectrum will eventually be shaped according to the desired speech spectrum. This shaping entails a masking effect which would further suppress the noise, from a perceptual standpoint, with a relatively lower signal distortion.

This suggests that adding further masking criteria, based on known human auditory properties, to the SSA suppression filters could result in improved noise shaping hence enhancing the overall performance. This is what we will try to achieve by the novel PSS method developed in this thesis and described in the next chapter.

4.3 The effect of noise

In the previous section we concluded that the SSA can be viewed from a filterbank standpoint where the analysis filters are data dependent. The passbands of those filters, especially the first few eigenvectors with largest eigenvalues, usually track the formant locations of the input speech signal. This conclusion, however, is based on clean signal covariance matrices. Therefore it would be informative if we can observe the behaviour of the analysis filters in the presence of noise.

To this end, let $V_{s,i}(\omega)$ be the DTFT of the i^{th} eigenvector calculated from a clean speech covariance matrix and let $V_{x,i}(\omega)$ be the one calculated from the noisy speech covariance matrix. Recall that the indices, i, are obtained by sorting the eigenvalues in decreasing order.

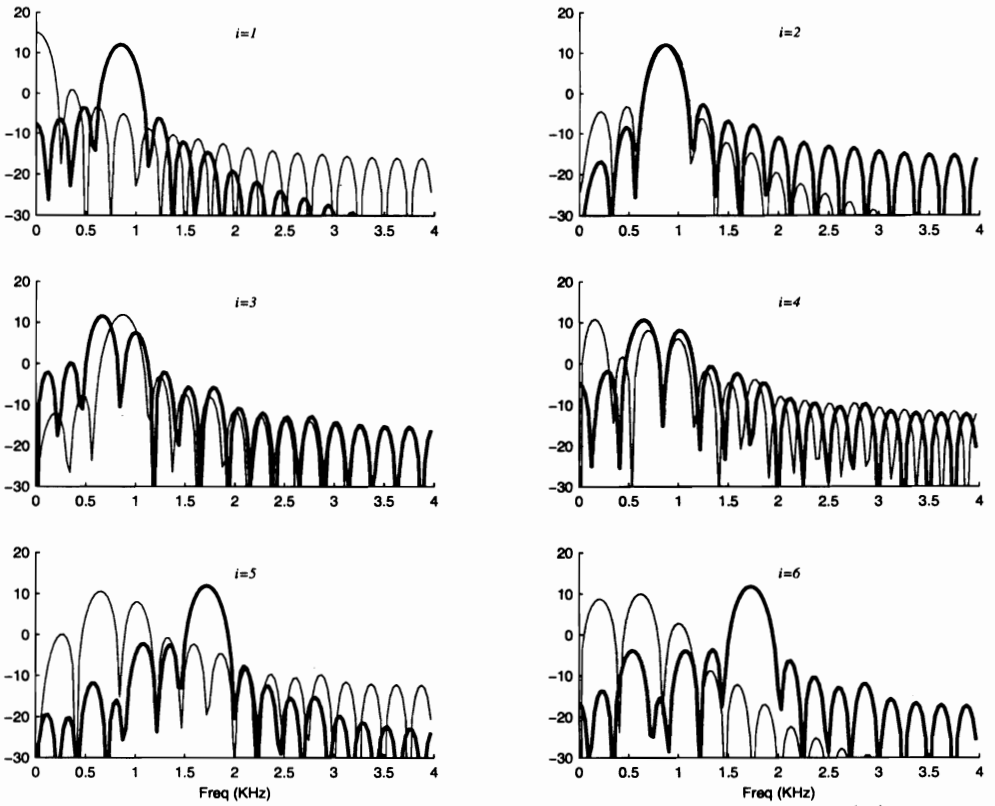


Fig. 4.4 The magnitude squared of the first six eigenfilters $V_i(\omega)$ for the vowel /a/ in the word "cats", calculated from a clean signal (thick) and from a noisy signal at 0 dB SNR (thin).

4.3.1 Using the noisy covariance matrix

Figure 4.4 shows a plot of $V_{s,i}(\omega)$ (thick line) and $V_{x,i}(\omega)$ (thin line) for $i = 1, \ldots, 6$, for the vowel /a/. The noise is that of a Volvo car added at 0 dB SNR. Figure 4.5 shows the corresponding spectra for the affricate /ch/.

It can be seen that in both cases, the analysis filters were significantly affected by the noise. Indeed, while $V_{s,1}(\omega)$, for example, had captured the speech first formant, the noise had shifted the passband of $V_{x,1}(\omega)$ to a lower frequency band. Actually, as can be seen in Figure 3.6, the noise, which has a low-pass nature, has most of its energy concentrated in this low frequency band. Therefore, this spectral peak caused by the noise, had been "treated" by the EVD as if it was the first speech formant[3]. The actual formant is only

[3]At 0 dB the noise energy was high enough for its corresponding eigen analysis filter to be placed in the first position by the sorting operation.

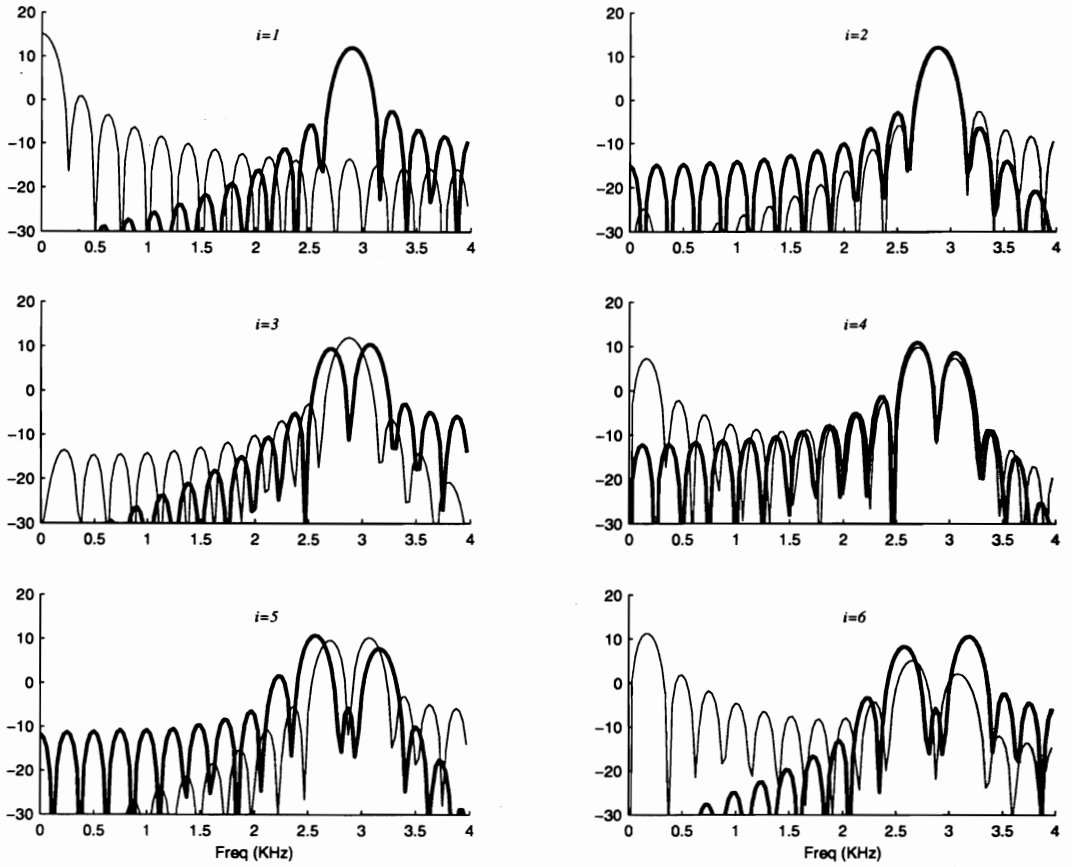


Fig. 4.5 The magnitude squared of the first six eigenfilters $V_{s,i}(\omega)$ (thick) and $V_{x,i}(\omega)$ (thin) for the affricate /ch/ under Volvo car noise at 0 dB SNR.

captured by the second filter as can be seen in Figure 4.4. The second formant, originally found in the passband of the fifth and sixth filters in Figure 4.2, has yet to be resolved. This result is expected since at 0 dB, the speech formants have less energy than that of noise and the frequency bands with the highest energy are no longer those of the formants. Indeed, our tests show that for SNR values above 5 dB, the passbands of $V_{x,1}(\omega)$ are almost identical to those of $V_{s,1}(\omega)$.

In [130] the KLT was obtained from the EVD of the noisy signal and satisfactory results were reported. The above discussion actually supports to some extent this choice for low to moderate noise levels. Although the eigenvectors would be different from those of a clean signal, the speech formants would eventually be resolved by the analysis eigenfilters. However, for best performance, and to isolate the noise energy, it is desirable to use an estimate of the clean speech covariance matrix.

In their method also, Rezayee and Gazor [130] propose a VAD based on tracking the principal component (i.e. the largest eigenvalue) of the noisy covariance matrix. Based on the previous discussion, transitions from non-speech to speech segments may not necessarily be reflected in the energy of the principal component. Actually, among others, the following two scenarios are likely to occur. Under very low SNR conditions, the first eigenvector (which corresponds to the principal component) would continue to track the frequency band of the noise energy peak whereas the speech formants will be resolved by other eigenvectors. Hence the presence of speech can pass undetected by the VAD. For example, this would be the case for weak fricatives or affricates (like /ch/ as shown in Figure 4.5), which are very likely to occur at word beginnings. Under high SNR conditions, the eigenfilter passband can shift to the location of the first formant hence neglecting (filtering out) the noise energy. Therefore, while the "origin" of the principal component energy had changed, it would not necessarily show an increase which would trigger the detection of speech presence.

The above described problem can result in inaccurate speech endpoints hence the cancellation of weak, though important, speech sounds. This is perceived as signal clipping at word boundaries. This artifact had been indeed reported in [130] though it is claimed that listeners did not consider this as a serious problem that hinders intelligibility.

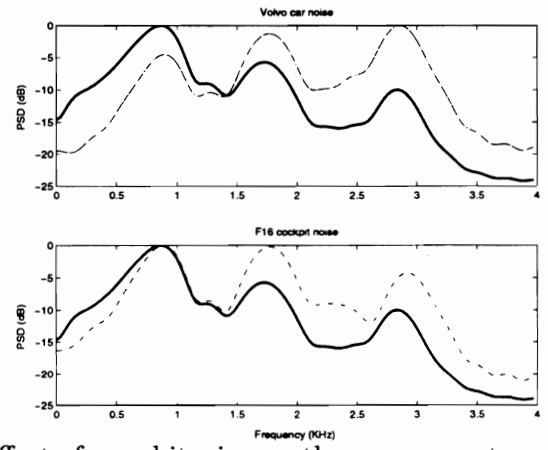


Fig. 4.6 The effect of prewhitening on the power spectrum of the vowel /a/, for two types of noise, Volvo car (up) and F16 jet cockpit (down): Original (thick) and after prewhitening (dashed).

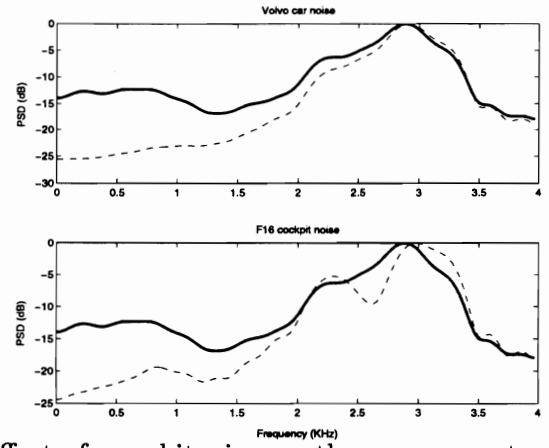


Fig. 4.7 The effect of prewhitening on the power spectrum of the affricate /ch/, for two types of noise, Volvo car (up) and F16 jet cockpit (down): Original (thick) and after prewhitening (dashed).

4.3.2 The effect of prewhitening

In the original SSA method [41], prewhitening was used to handle colored noise. As explained in Section 3.4.5, the input noisy vector $\mathbf{x}$ is prewhitened by multiplying it with $\mathbf{R}_w^{-1/2}$, where $\mathbf{R}_w$ is the noise covariance matrix. Then the EVD of the covariance matrix $\mathbf{R}_{\check{x}}$ of the prewhitened signal, $\check{\mathbf{x}} = \mathbf{R}_w^{-1/2}\mathbf{x}$, is used to design the signal subspace filter. Let the analysis eigen filter corresponding to the i^{th} eigenvector of $\mathbf{R}_{\check{x}}$ be denoted as $V_{pw,i}(\omega)$.

The effect of this prewhitening is shown in Figures 4.6 and 4.7 for /a/ and /ch/ respectively. It can be seen that while the position of the formants remains unchanged, their energies, relative to each other, do change. This can be seen in the case of /a/ where the three formants after prewhitening, have almost the same energy level, with the third formant becoming the dominant one in the case of the Volvo car noise. The speech spectra of the affricate /ch/, on the other hand, appear to be less affected by prewhitening. This can be explained by the fact that this particular phoneme has a unique formant occurring in the higher bands of the frequency range of interest. Hence, the spectrum is less vulnerable to the prewhitening matrix corresponding to the low pass Volvo car noise. With the F16 noise, the high frequency peak in the noise spectrum (as can be seen in Figure 3.7) inflicted a larger effect where the wide formant has been split into two slightly separated peaks. These modification made to the speech spectra, affect the perceptual information they carry. For this reason, and since this information is important to calculate a masking threshold, we rule out the option of prewhitening while designing the perceptual signal

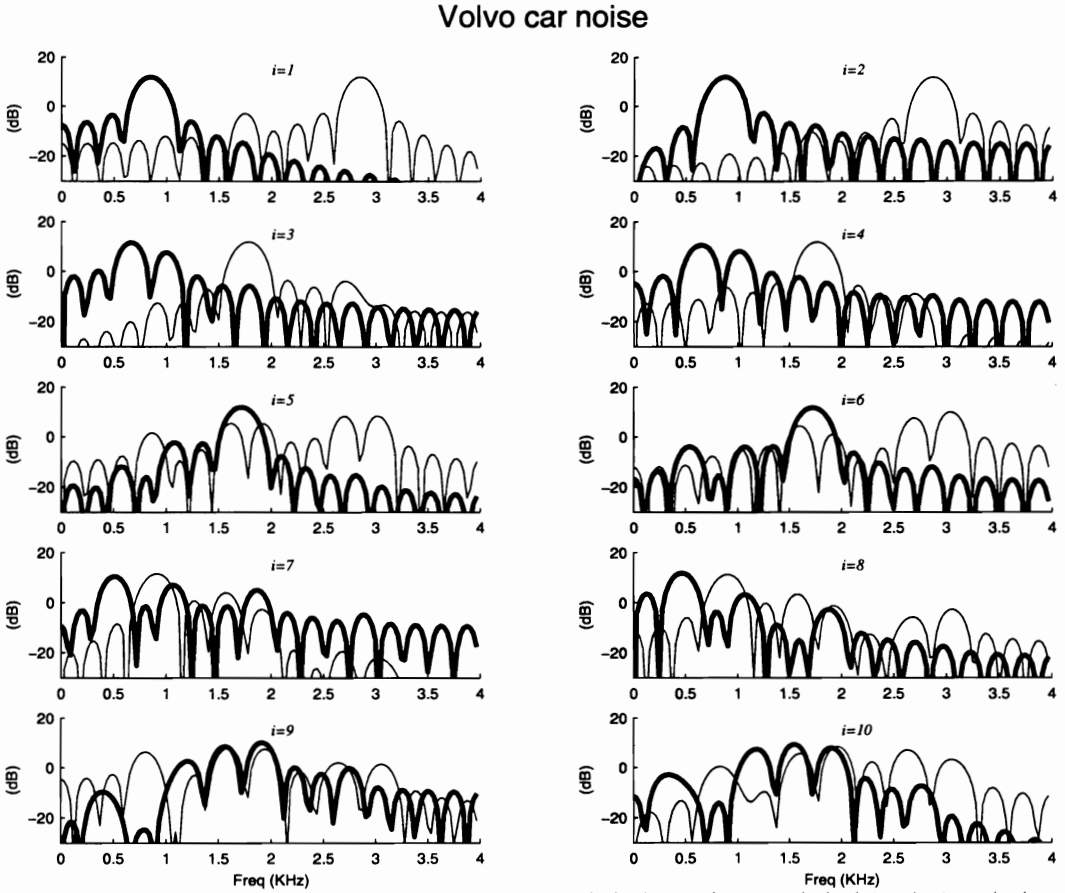


Fig. 4.8 The magnitude squared of $V_{s,i}(\omega)$ (thick) $V_{pw,i}(\omega)$ (thin) for /a/ under Volvo car noise at 0 dB SNR.

subspace method.

The effect of prewhitening can also be understood via the eigen analysis filters $V_{pw,i}(\omega)$. Indeed, the dominant formant in the prewhitened PSD of /a/ (when corrupted by Volvo car noise) became the third formant. Therefore, as can be seen in Figure 4.8, the passbands of $V_{pw,1}(\omega)$ and $V_{pw,2}(\omega)$ are now located at the frequency band of of the third formant instead of the first formant as it is the case with $V_{s,1}(\omega)$ and $V_{s,2}(\omega)$. For the F16 cockpit noise, the first and second formants of the prewhitened signal have the highest energies hence it can be seen from Figure 4.9 that the passband of $V_{pw,1}(\omega)$ and $V_{pw,2}(\omega)$ are spread over the frequency bands occupied by these two formants. In both noise type cases, the most important first formant has not been uniquely isolated by any of the first ten eigen analysis filters shown in the Figures.

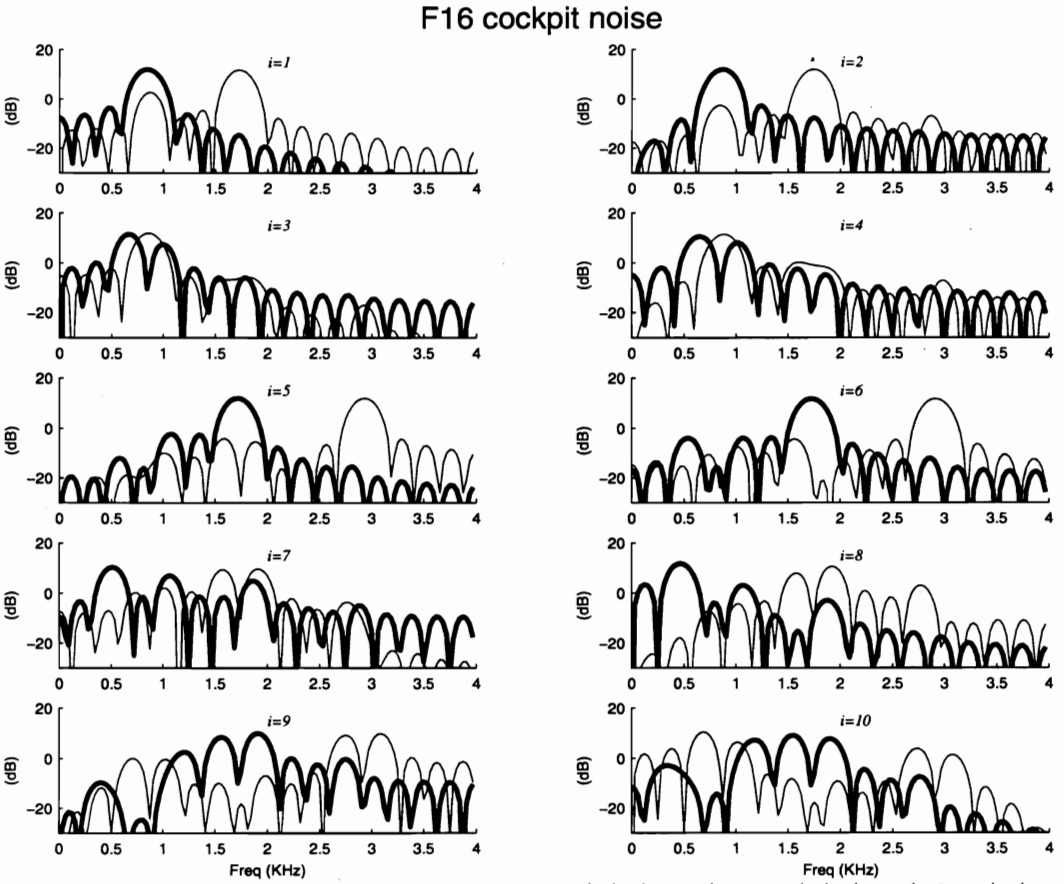


Fig. 4.9 The magnitude squared of $V_{s,i}(\omega)$ (thick) $V_{pw,i}(\omega)$ (thin) for /a/ under F16 cockpit noise at 0 dB SNR.

The affricate /ch/, on the other hand, and due to its spectral characteristics, is less affected by prewhitening as can be seen in Figure 4.7. Therefore, $V_{pw,i}(\omega)$'s are closer to $V_{s,i}(\omega)$'s with a slight shifting effect with F16 noise (Figure 4.11) and an almost exact match with the Volvo noise (Figure 4.10) especially for the first few eigen analysis filters.

4.4 Properties of the Blackman-Tukey Spectrum estimator

Since the Inverse FET provides a PSD estimate based on the Blackman-Tukey spectrum estimator, we found it necessary to examine the properties of this estimator to verify how adequate it is for the current application.

The periodogram is a very popular spectrum estimator because it can be directly calculated from the samples of $x(n)$, as shown in (3.15). However its drawback is that it suffers

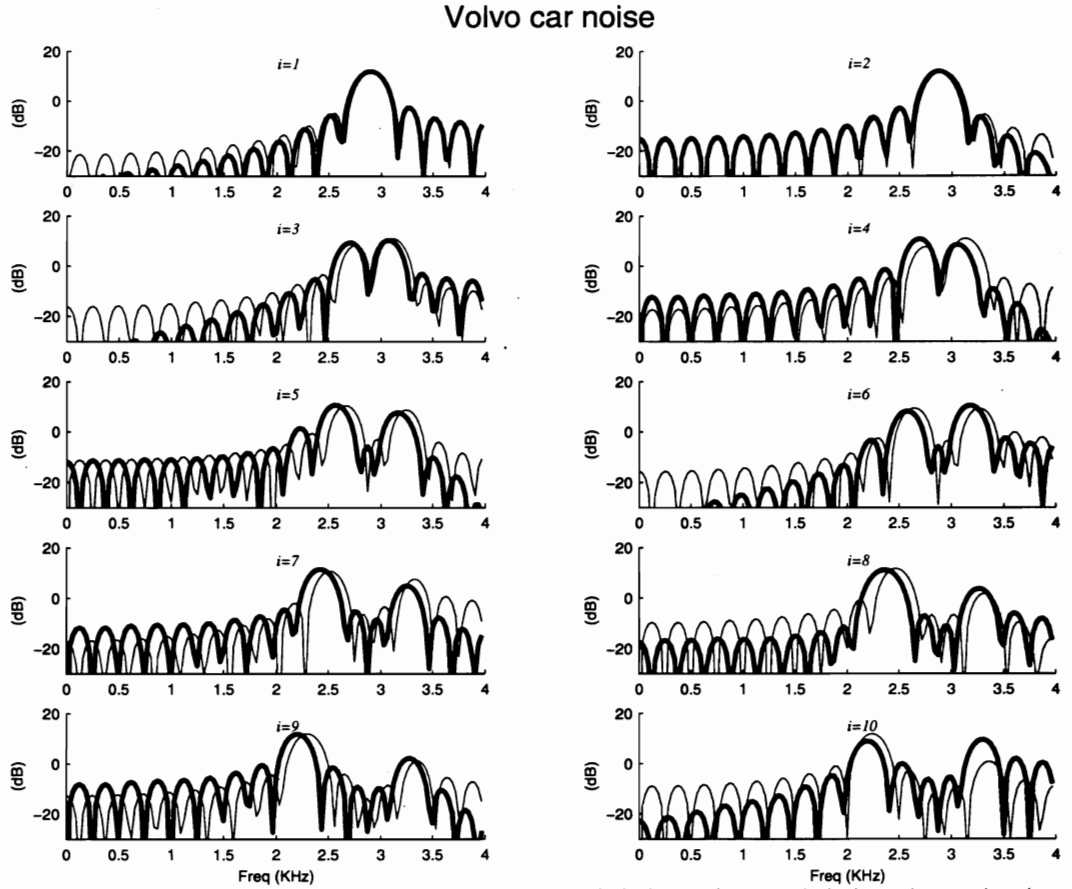


Fig. 4.10 The magnitude squared of $V_{s,i}(\omega)$ (thick) $V_{pw,i}(\omega)$ (thin) for /ch/ under Volvo car noise at 0 dB SNR.

from a high variance given by [94],

$$\mathrm{Var}\{\Phi(\omega)\} \approx \tilde{\Phi}^2(\omega) \tag{4.19}$$

where $\tilde{\Phi}(\omega)$ is the exact PSD of the underlying random process. This variance is in general considered to be high and can not be tolerated. Precisely, in the current application, the same signal subspace filter designed using the FET, will be applied to several overlapping adjacent vectors as will be discussed in section 5.1. Therefore, it is preferable that the designed filter have a minimal variance.

In the Blackman-Tukey estimator, the variance is reduced by multiplying the auto-correlation function by the window. The variance in the case of a Bartlett window is

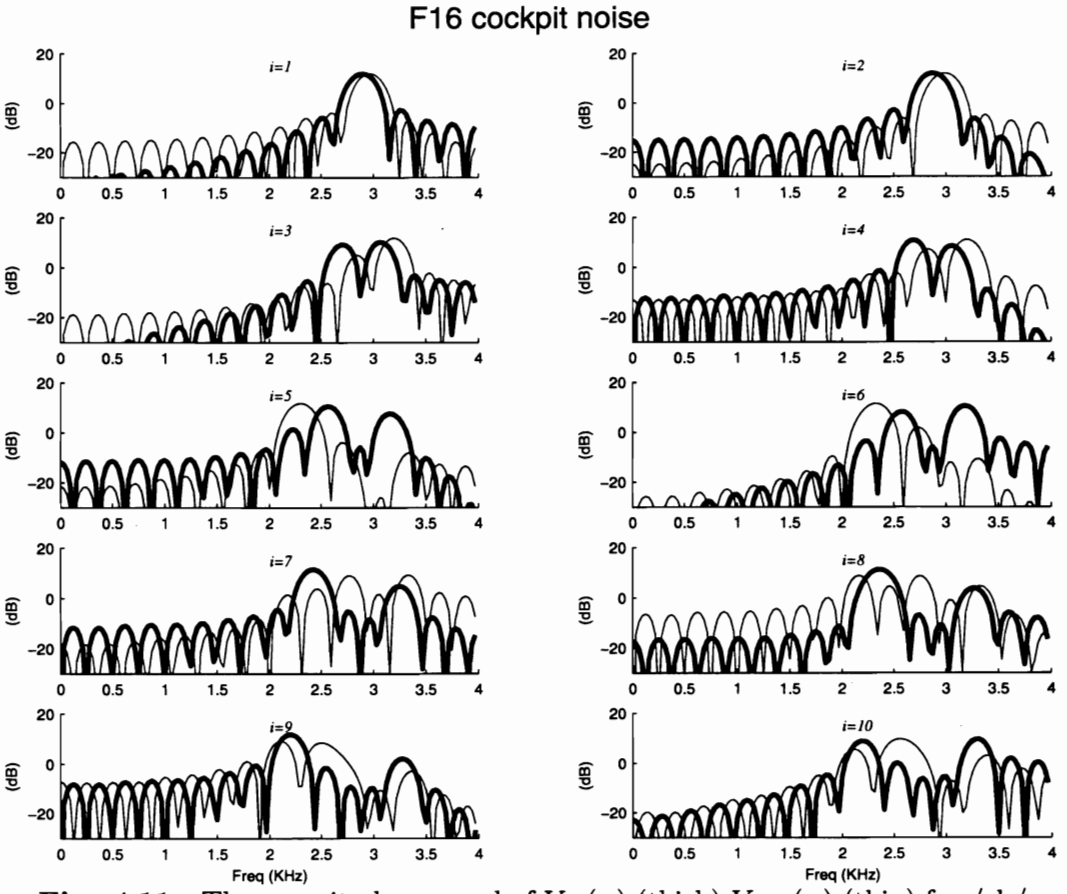


Fig. 4.11 The magnitude squared of $V_{s,i}(\omega)$ (thick) $V_{pw,i}(\omega)$ (thin) for /ch/ under F16 cockpit noise at 0 dB SNR.

approximately [94]

$$\mathrm{Var}\{\Phi_B(\omega)\} \approx \tilde{\Phi}^2(\omega)\frac{1}{L}\sum_{i=-P}^{P} w_b^2(i) \approx \tilde{\Phi}^2(\omega)\frac{2P}{3L} \tag{4.20}$$

which is less than $\mathrm{Var}\{\Phi(\omega)\}$ since $P \leq L$.

This lower variance is obtained at the expense of a reduced resolution. The Blackman-Tukey estimate is a smoothed version of the periodogram due to the convolution with the Fourier Transform of the window in the frequency domain. So the resolution depends on the bandwidth of the main lobe of the window which in turn depends on its size and type.

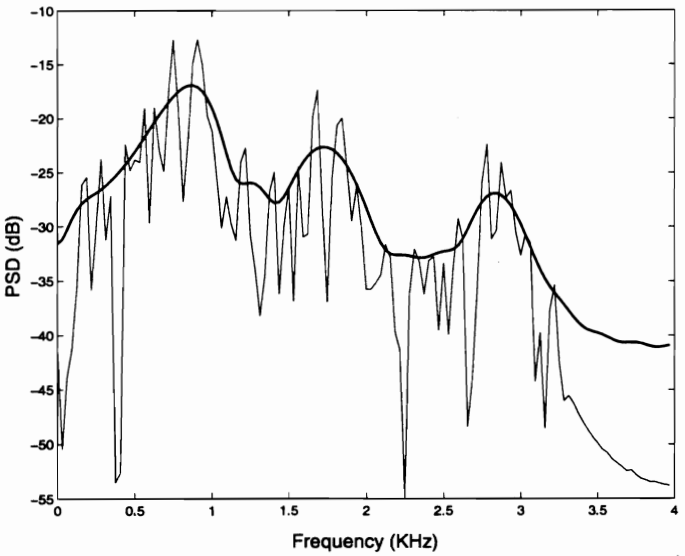


Fig. 4.12 The power spectral density estimate for the vowel /a/ obtained using the periodogram (thin) and the Blackman-Tukey estimate (thick).

In our case, for a length $2P-1$ Bartlett window the resolution $\Delta\omega$ is given by [64]

$$\Delta\omega \approx 0.64\frac{2\pi}{P} \tag{4.21}$$

So for $P = 32$ at 8 KHz sampling rate, the resolution will be 160 Hz which will result in a wideband spectrum which smoothes the fine structure of the harmonics while preserving formant structure. For example in the case of vowels, the first three formants, important for speech intelligibility, are on the average 1 KHz apart [124] so they will be well identified with the Blackman-Tukey spectral estimator.

The resolution of the periodogram, on the other hand, is $0.89\frac{2\pi}{L}$, that is 28 Hz when $L = 256$ [64]. So the periodogram will reveal unnecessary details for the present application where the goal is to calculate the masking threshold which is a smooth function of frequency due to the linear and nonlinear transformations applied to the input speech signal PSD.

In Figures 4.12 and 4.13, the periodogram and the Blackman-Tukey estimates of the vowel /a/ and the affricate /ch/ respectively, are shown. It can be seen that the formants can be clearly seen in the BT estimate whose low variance is also evident. The periodogram, on the other hand, reveals a high frequency resolution coupled with a high variance. Indeed this high variance is behind the fluctuations in the calculated filter coefficients, in say spectral subtraction, which causes the creation of the musical noise artifact [38].

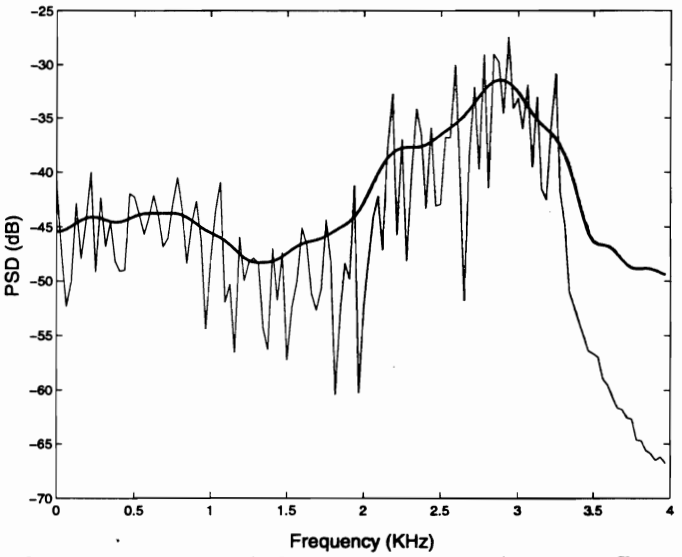


Fig. 4.13 The power spectral density estimate for the affricate /ch/ obtained using the periodogram (thin) and the Blackman-Tukey estimate (thick).

4.5 Implementation

If the FET and IFET relationships (4.4) and (4.11) respectively, are to be useful in practice, an implementation using the Discrete Fourier Transform (DFT) instead of the DTFT should be provided. This would allow the use of a matrix-vector product operation, which can be easily handled by digital computers .

In practice we have available a signal $x(n)$ for which the covariance matrix estimate $\mathbf{R}$ is calculated. The EVD of $\mathbf{R}$ is obtained as $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_P]$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_P)$. The eigenvalues can also be expressed as a vector in the following way

$$\boldsymbol{\lambda} = [\lambda_1,\ \lambda_2,\ \ldots\ , \lambda_P]^T \tag{4.22}$$

Consider now any eigenvector $\mathbf{u}_i = [u_i(0), \ldots, u_i(P-1)]^T$. As mentioned earlier, this vector can be viewed as a length-P signal $u_i(p)$ which is equal to zero outside the interval $[0, P-1]$. The K-point DFT of $u_i(p)$, defined in (3.7), is given by

$$V_i(k) = \sum_{p=0}^{P-1} u_i(p) e^{-j2\pi kp/K} \quad \text{for } k = 0, \ldots, K-1 \tag{4.23}$$

where $K \geq P$ is assumed. Using (3.8), $V_i(k)$ is related to the DTFT $V_i(\omega)$, defined in (4.5),

by

$$V_i(k) = V_i(\omega)\Big|_{\omega=\frac{2\pi}{K}k} \quad \text{for} \quad k = 0, \ldots, K-1 \tag{4.24}$$

Now define the vectors

$$\mathbf{v}_i = [v_i(0), \ldots, v_i(K-1)]^T \quad \text{for} \quad i = 1, \ldots, P. \tag{4.25}$$

where

$$v_i(k) = |V_i(k)|^2 \quad \text{for} \quad k = 0, \ldots, K-1 \tag{4.26}$$

and let the $K \times 1$ vector $\mathbf{\Phi}_B = [\Phi_B(0), \ldots, \Phi_B(K-1)]^T$ be the discrete Blackman-Tukey estimate such that

$$\Phi_B(k) = \Phi_B(\omega)\Big|_{\omega=\frac{2\pi}{K}k} \quad \text{for} \quad k = 0, \ldots, K-1$$

$\mathbf{\Phi}_B$ can be written using the IFET relationship (4.11) in the following way

$$\mathbf{\Phi}_B = \frac{1}{P}\sum_{i=1}^{P} \lambda_i \mathbf{v}_i \tag{4.27}$$

In matrix notation, (4.27) can readily be written as

$$\mathbf{\Phi}_B = \frac{1}{P}\mathbf{V}\boldsymbol{\lambda} \tag{4.28}$$

where $\mathbf{V}$ is a $K \times P$ matrix given by

$$\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_P] \tag{4.29}$$

Since $\lambda_i = 0$ for $i > Q$, it is enough to just retain the first Q columns so that $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_Q]$. For the same reason, in practice, we have $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_Q]^T$.

In a similar way, if a PSD estimate $\mathbf{\Phi} = [\Phi(0), \ldots, \Phi(K-1)]^T$ is available, a set of energies $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_P]^T$ reflecting the spectral content of $\mathbf{\Phi}$ into the eigen-domain can be obtained using the FET relationship (4.4) which can be approximated with the following implementation

$$\boldsymbol{\xi} = \frac{1}{K}\mathbf{V}^T\mathbf{\Phi} \tag{4.30}$$

4.5.1 Selecting the DFT size

As can be noted, equations (4.30) and (4.28) are an approximation to the FET and IFET relationships respectively, obtained via the DFT. Therefore, to assess the validity of this approximation, one needs to find the condition on the value of the DFT size.

Using the definitions given earlier, (4.30) can be written as

$$\xi_i = \frac{1}{K}\sum_{k=0}^{K-1}\Phi(k)|V_i(k)|^2 \tag{4.31}$$

where $|V_i(k)|^2$ is given by (4.24) and is related to the length P sequence $u_i(p)$ by

$$V_i(k) = \sum_{p=0}^{P-1} u_i(p)e^{-\frac{2\pi}{K}pk} \tag{4.32}$$

Substituting $V_i(k)$ in (4.31) and rearranging the summations we obtain

$$\xi_i = \sum_{p=0}^{P-1}\sum_{q=0}^{P-1} u_i^*(p)u_i(q)\underbrace{\frac{1}{K}\sum_{k=0}^{K-1}\Phi(k)e^{\frac{2\pi}{K}(p-q)k}}_{\hat{r}(p-q)} \tag{4.33}$$

Comparing the above equation with (4.6), repeated below for convenience

$$\lambda_i = \sum_{p=0}^{P-1}\sum_{q=0}^{P-1} u_i^*(p)r(p-q)u_i(q) \tag{4.34}$$

it can be concluded that $\xi_i = \lambda_i$ if $\hat{r}(p) = r(p)$ where

$$\hat{r}(p) = \frac{1}{K}\sum_{k=0}^{K-1}\Phi(k)e^{\frac{2\pi}{K}pk} \tag{4.35}$$

Now suppose that $x(n)$ is a length-L sequence, i.e. $x(n) = 0$ for $n < 0$ and $n > L-1$ and that $\Phi(k)$ is obtained as the periodogram of $x(n)$ given by

$$\Phi(k) = \frac{1}{L}|X(k)|^2 = \frac{1}{L}X(k)X^*(k) \tag{4.36}$$

where $X(k)$ is the K-point DFT of $x(n)$. Then, (4.35) can be written as

$$\hat{r}(p) = \frac{1}{L}\frac{1}{K}\sum_{k=0}^{K-1} X(k)X^*(k)e^{\frac{2\pi}{K}pk} \tag{4.37}$$

Recognizing $X^*(k)$ as the DFT of $x(-n)$, $\hat{r}(p)$ is actually the (scaled) circular convolution of $x(n)$ and $x(-n)$, i.e. a circular autocorrelation [123]. Therefore to force this circular autocorrelation to be equal to the desired autocorrelation $r(p)$, over the interval $0 \leq p \leq P-1$, the DFT should satisfy the condition[4] $K \geq L + P - 1$ [123].

On the other hand, suppose that the signal PSD $\Phi(k)$, was the Blackman-Tukey estimate defined as

$$\Phi_B(k) = \sum_{q=-P+1}^{P-1} r(q)w(q)e^{-\frac{2\pi}{K}qk} \tag{4.38}$$

where $w(q)$ is a length $2P-1$ window. In this case the best that can be achieved is

$$\hat{r}(p) = r(p)w(p) \quad \text{for } -P+1 \leq p \leq P-1 \tag{4.39}$$

where now the condition required is $K \geq 2P-1$. Actually this condition should be satisfied in order to obtain the BT estimate using the IFET implementation (4.28).

Therefore, for $P = 32$, the FET and IFET relationships can be implemented using a DFT of size $K = 64$. However, and since the obtained BT spectrum is used to calculate a masking threshold and then mapped again to the eigendomain, a larger DFT size was experimentally found to yield better results. This is because the used masking model gives better masking threshold estimates if more frequency lines are used. For this reason the DFT size is set to $K = 256$, which equals the frame length considered in this thesis.

4.6 Summary

In this chapter we presented a Frequency to Eigendomain Transformation (FET). It basically sets a bridge between the frequency domain and the eigendomain. The FET served as an analysis tool which offered a different interpretation of the SSA. Namely, a filter-bank interpretation with the eigenvectors as analysis filters and the eigenvalues as the total

[4] Actually this condition is required when calculating the autocorrelation sequence efficiently via the FFT [123].

energy at the output of those filters. The analysis provided supports the claim that the SSA results in a better residual noise spectrum shaping (hence less annoying musical noise) than, say, spectral subtraction. This is because the KLT makes the suppression focused around the frequency bands occupied by the speech formants.

Using the FET we also attempted to explain the results reported in the literature for some SSA based methods. The benefit of this is to acquire a more comprehensive understanding of the SSA allowing to avoid its drawbacks and to take the most advantage of its strong points. The analysis performed here shall also motivate some of our design decisions which are to be included in the new perceptual signal subspace method we present in the next chapter.

Chapter 5

The Perceptual Signal Subspace method

In this chapter we seek to further improve the noise shaping capabilities of the SSA by altering its noise reduction mechanism according to perceptual criteria. This can be achieved by modifying the gain function so that instead of being dependent on the SNR along a particular direction (or subband), it becomes dependent on the ratio of the masking energy to the noise energy along that direction. This approach is driven by the assumption that as long as the noise is below a masking threshold, it will not be audible. Therefore, recalling the inevitable trade-off between the extent of noise reduction and the resulting signal distortion, it would be beneficial if we avoid to suppress any noise component which is not perceived any way.

In addition to its benefit as an analysis tool, the FET can be used as a design tool which permits to calculate the masking energy along every spectral direction, via a sophisticated masking model. The resulting method is called the Perceptual Signal Subspace method (PSS) which will be experimentally shown, in Chapter 8, to provide a better noise reduction performance than the classical SSA by offering a better trade off between the signal distortion and the residual noise level. The PSS is designed so that it can handle the general case of colored noise.

The main handicap of the SSA remains its relatively high computational load resulting from the expensive EVD computation. In this chapter we also show how this burden can be considerably reduced without any performance side-effects. The idea simply exploits the

stationarity assumption of the speech signal in the same way it is exploited by traditional frequency domain methods. The proposed new technique can be used with any SSA based method as an alternative to the original implementation scheme.

5.1 Calculating the eigenvalue decomposition

The disadvantage of the signal subspace approach is the relatively high computational load mainly due to the expensive eigenvalue decomposition. This drawback made the engineers working on speech enhancement rather reluctant to use the SSA in practice. However, with the impressive development in the DSP technology and the continuous increase in the available processing speed and computational power, it is believed that the more robust SSA can eventually compete with the widely used frequency domain methods.

The complexity issue has however been addressed in the literature and several approaches have been proposed to tackle this problem. These techniques include fast EVD and subspace tracking methods. Moreover, attempts to approximate the KLT using the Discrete-Cosine Transform (DCT) have also been recently applied to speech enhancement [71, 156].

We here briefly describe some of these methods then we propose a novel technique developed in this thesis which reduces the computational load without any side-effect on the noise reduction performance of the SSA.

5.1.1 Fast eigenvalue decomposition methods

One solution for the complexity issue is to replace the exact EVD by other fast subspace methods which are capable to reduce the complexity from $\mathcal{O}(P^3)$ to $\mathcal{O}(P^2Q)$ operations per sample where Q is the rank of the matrix. For example in [166], the structure of the covariance matrix is exploited and the so-called Lanczos algorithm [50] is used to reduce the required computations by only calculating the Q principal eigenvectors (and their corresponding eigenvalues) which span the signal subspace. For example, for a speech signal sampled at 8 KHz, the effective rank Q of the covariance matrix of a voiced frame (which constitute the majority of a speech sentence) would be around 10 to 15, for a $P = 32$ model. Therefore, computational savings could be achieved by approximating the exact EVD using such fast methods.

5.1.2 Subspace tracking methods

Another technique which follows a similar scheme is subspace tracking. Rather than trying to calculate the EVD from scratch, the subspace trackers seek to update an already existing EVD as more data becomes available. In [130] a fast implementation of the SSA for speech enhancement has been developed based on the Projection Approximation Subspace Tracking (with deflation) method (PASTd) which reduces the complexity to $\mathcal{O}(PQ)$ per sample using the Recursive Least Squares (RLS) algorithm [167]. The PASTd algorithm does not guarantee the orthogonality of the eigenvectors, this is why it was reported in [85] that better results can be achieved using the Fast Orthogonal Iteration (FOI) based algorithm [145]. In [62] a rank-revealing ULV decomposition [144] has been also used for speech enhancement. Another $\mathcal{O}(PQ)$ subspace tracking algorithm based on Givens rotations and which guarantees the orthogonality of the eigenvectors is also reported in [19] but has yet to be tested in a speech enhancement application.

Unfortunately, there seem to be some problems associated with applying subspace trackers to speech enhancement. The reason is that these methods are based on estimating the covariance matrix using a sliding exponential window. During our experiments however, we noticed that shifting the window at a high rate added reverberation to the enhanced speech signal. This result, which has also been confirmed in [85], suggests that a sliding exponential window may be inadequate for speech enhancement applications. Therefore, subspace trackers can only be applied if the EVD update scheme is carried out on a sample by sample basis[1], which does not lead to the apparent great computational savings. In fact, an exact EVD has a computational cost of $\mathcal{O}(P^3)$, but since it is only calculated every $P/2$ samples, this complexity is reduced to $\mathcal{O}(P^2)$ per sample[2]. A subspace tracker on the other hand can at best achieve a $\mathcal{O}(PQ)$ per sample complexity.

5.2 The Frame Based EVD (FBEVD) method

In this thesis we develop a novel implementation scheme which helps to overcome the computational issue of the SSA. The method we propose here is a modification of the approach used in [41] and described in Section 3.4.4. The idea is based on the stationarity

[1] As in the case of the method reported in [130].

[2] Moreover, if a fast eigenvalue decomposition is used, for example [166], the complexity would be $\mathcal{O}(PQ)$ per sample.

assumption of the speech signal. This assumption is already exploited in [41] to calculate the covariance matrix required for the subspace filter design, but we apply it here in a different manner.

5.2.1 Description

Let the speech signal $x(n)$ be divided into length L overlapping frames $x_i(m)$ with a shift of D samples,

$$x_i(m) = x(iD + m), \quad m = 0, \ldots, L-1 \tag{5.1}$$

This frame is used to obtain the biased autocorrelation function estimate as follows

$$r_x(p, i) = \frac{1}{L} \sum_{m=0}^{L-p} x_i(m) x_i(m+p) \quad p = 0, \ldots, P-1 \tag{5.2}$$

These autocorrelation coefficients are used as described in Section 3.4.4 to calculate the subspace filter $\mathbf{H}_i$. Note that this filter has now a subscript i to emphasize the fact that it is computed based on the signal samples of the i^{th} frame.

Every frame is divided into smaller P-dimensional overlapping vectors with a 50% overlap as shown if Figure 5.1. The frame length L is chosen to be a multiple of P so that there will be in total $\frac{2L}{P} - 1$ vectors in one frame. Like all frequency domain methods, the speech signal within every frame is assumed to be stationary so that these vectors would all have the same covariance matrix and hence the same subspace filter $\mathbf{H}_i$. Therefore we have

$$\hat{\mathbf{s}}_{i,jP/2} = \mathbf{H}_i \mathbf{x}_i(jP/2) \quad \text{for } j = 2, \ldots, \frac{2L}{P} \tag{5.3}$$

where the input vector $\mathbf{x}_i(m) = [x_i(m), x_i(m-1), \ldots, x_i(m-P+1)]^T$ and the filter output $\hat{\mathbf{s}}_{i,m}$ is defined in a similar way. The output vectors are then multiplied by a length-P Hanning window and synthesized using the overlap-add method to obtain one enhanced frame $\hat{s}_i(m) = \hat{s}(iD + m)$. Finally every frame is multiplied by a second length-L Hanning window and the total enhanced speech signal is recovered using the overlap-add synthesis technique. A 50% overlap is also applied to these larger analysis frames, that is $D = L/2$. In this way every input vector is enhanced using filters designed from two different analysis frames which allows to compensate for any speech non-stationarity.

This frame based approach is analogous to frequency domain methods where frame

overlap is applied, with every length-L frame of noisy speech being enhanced using a unique filter.

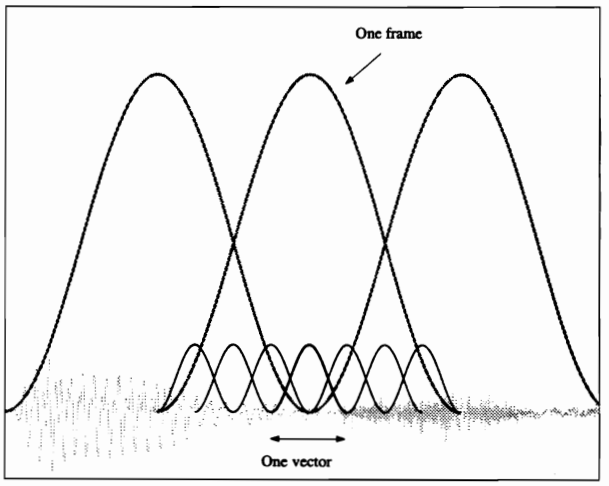


Fig. 5.1 Illustration of the partition of the speech signal into frames and vectors.

5.2.2 Computational savings

In the original SSA implementation described in Section 3.4.4 the EVD, with complexity $\mathcal{O}(P^3)$, is carried out every $P/2$ samples resulting in a total complexity in the order of $\mathcal{O}(P^2)$ per sample as discussed earlier.

In the new FBEVD scheme, the EVD is only needed every frame at a rate of $L/2$ samples, where L is the frame length. Thus, if $L = \kappa P$ then the computational cost of the EVD would reduce to $\mathcal{O}(P^2/\kappa)$. That is the computational savings will be proportional to $\kappa = L/P$.

For example for $L = 256$ and $P = 32$, we have $\kappa = 8$. This results in reducing the cost of EVD calculation by a factor of 8. Knowing that the largest computational burden of the SSA arises from the EVD, this factor constitutes a significant saving at almost no performance degradation as will be shown in the experimental results. Coupling this method with one of the fast EVD techniques discussed earlier would considerably reduce the overall computational load.

The FBEVD will be evaluated in Chapter 8 where the computational savings and the incurred performance degradation, if any, will be measured.

5.3 The perceptual gain function

As discussed in Section 3.4, the SSA consists of projecting the input noisy speech vector onto the signal subspace and then suppressing any remaining noise by multiplying the signal energy along every spectral direction by a specific gain. This gain is chosen to be a function of the signal to noise ratio where this function satisfies condition (3.71).

The objective is to apply severe noise suppression along a spectral direction if the corresponding SNR is low. This SNR, defined in Section 3.4.2, is given here again for convenience

$$\gamma_i = \frac{\lambda_{s,i}}{\xi_i} \quad \text{for } i = 1, \ldots, Q \tag{5.4}$$

where ξ_i is the noise energy along the i^{th} direction as defined in (3.91) and $\lambda_{s,i}$ is the i^{th} eigenvalue of the clean speech signal.

Since the noise becomes inaudible if it is below the masking threshold, it is advantageous to modify the gain function making it dependent on a perceptually significant quantity. This quantity, hereafter referred to as the Mask to Noise Ratio (MNR), is defined as

$$\bar{\gamma}_i = \frac{\theta_i}{\xi_i} \quad \text{for } i = 1, \ldots, Q \tag{5.5}$$

that is, it is the ratio of the masking energy θ_i to the noise energy ξ_i for the i^{th} spectral component.

From a filterbank perspective, the above quantity provides a measure of the audible noise at the output of every analysis filter. Note that this approach is analogous to the one adopted by Tsoukalas [150] and Gustafsson [56] as shown in (3.34) and (3.37) respectively. In those frequency domain methods, different gain functions are used but both are dependent on the MNR which is calculated for every frequency[3].

The MNR (5.5) can now be used for the gain function $f(\bar{\gamma}_i)$. As was mentioned earlier, the exponential gain function (3.74) is being used in this thesis. Hence the diagonal entries of the gain matrix will be given by

$$g_i = e^{-\nu \xi_i / \theta_i} \tag{5.6}$$

During our experiments we observed that in most cases we have $\theta_i < \lambda_{s,i}$ so that $\bar{\gamma}_i < \gamma_i$.

[3] In [150] the MNR is actually kept constant within one critical band as discussed in Section 3.3.2.

Therefore using the gain function (5.6) results in a more severe noise cancelling (recall that the gain function is an increasing function of γ_i as can be seen in Figure 3.2). However since the gain is now obtained via a perceptual criterion, the control parameter ν can be reduced in order to obtain less signal distortion without making the residual noise more audible. Our experiments show that an acceptable range is $0.5 \leq \nu \leq 1$ with $\nu = 0.8$ being a satisfactory value for most conditions. As ν increases beyond that value some undesired signal distortion starts to occur.

Nonetheless, during weak energy frames, such as unvoiced fricatives, the spectrum is rarely characterized in terms of formants because low frequencies are not excited and the excited upper resonances have broad bandwidths [124]. This can be verified in Figure 4.3 (or 4.7) for the affricate /ch/ as compared to the vowel /a/ in Figure 4.2 (or 4.6). Therefore, in such cases, the masking threshold estimate may not be accurate enough and it can happen that $\lambda_{s,i}$ be smaller than θ_i with the result that, if θ_i is used, not enough noise reduction is achieved, due to estimation errors. In particular, at transitions from silence to speech activity periods, the residual noise has a non smooth character which may be uncomfortable to some listeners.

Our informal listening tests show that modifying the gain function by taking the minimum of $\lambda_{s,i}$ and θ_i helps to improve the performance. The gain function hence becomes

$$g_i = e^{-\nu \xi_i / \min(\lambda_{s,i}, \theta_i)}. \tag{5.7}$$

This gain function is the one used in this thesis and the corresponding enhancement method is referred to as the Perceptual Signal Subspace (PSS) method.

Other gain functions

During our research some other gain functions have also been tested. Although the gain (5.7) was found to have the best performance, we found it beneficial to mention here the other options too.

The first alternative is based on the idea that if the MNR is greater than 1, along some spectral direction, then no noise suppression is required on that direction since the noise

would be masked anyway. Hence the gain function would be given by

$$g_i = \begin{cases} e^{-\nu\xi_i/\lambda_{s,i}} & \text{if } \xi_i > \theta_i \\ 1 & \text{else} \end{cases} \tag{5.8}$$

This gain function resulted in a slightly less distorted signal but failed to reduce the intensity of the musical noise.

Using the same reasoning as in (5.6), the above gain function can be modified as follows

$$g_i = \begin{cases} e^{-\nu\xi_i/\theta_i} & \text{if } \xi_i > \theta_i \\ 1 & \text{else} \end{cases} \tag{5.9}$$

In this case a slight improvement is achieved but the musical noise is still not as much masked as it is in the case of (5.7). Attempting to do so by increasing ν results in added signal distortion. It should be noted though that on some noise types such as car noise, this gain function gave better performance than RQSS.

The third and final alternative worth mentioning is to make the gain function dependent on the ratio of the signal energy to the difference between the masking and noise energies, that is

$$g_i = e^{-\nu \max(\xi_i - \theta_i, 0)/\lambda_{s,i}} \tag{5.10}$$

In this case, like in (5.8) and (5.9), the gain is set to one if the noise energy is below the masking energy. Otherwise, the amount by which the noise exceeds the masking energy is accounted for in the gain function (5.10). Interesting results have been achieved with this gain where the signal distortion had been considerably reduced. However, the residual noise had sometimes an annoying character (different from that of the musical noise) which cannot be tolerated by some listeners. For this reason (5.7) was preferred. It should be noted that for best performance, ν, in (5.10), should be set here to a higher value (around 7 or 8).

5.4 Calculating the masking threshold

In the course of our research, two masking models were used. Optimizing the masking model is in fact beyond the scope of this thesis but the use of a model which would provide the expected performance boost is essential. We first have used the Johnston model and

some initial experimental results were obtained. However, we found that better results are achieved with MPEG1 model 1 (hereafter the MPEG model) described in Section 2.3.2. For this reason, this model has been adopted in this thesis and unless otherwise mentioned the results reported here are based on it. Our choice was motivated, on one hand, by the satisfactory results we were able to achieve with this model, and on the other hand, by the wide spread and successful use of this model in audio coding applications. Our choice was also motivated by its relative implementation simplicity[4].

The steps required to calculated the masking threshold using the MPEG model are described in Section 2.3.2 and further details can be found in [12]. In our implementation we basically followed the same described steps except for labeling tonal and non-tonal components. Actually we needed to alter the labeling steps for the following reason.

In the original model, tonal components are selected based on finding local maxima and comparing their magnitude with some thresholds to omit the outliers. After finding the tonal components, all other frequencies are designated as non-tonal. However to reduce the computation, just a single non-tonal component is allowed within one critical band with its location chosen to be the geometric mean of the frequencies within that band. This results in a smooth threshold compared to the original periodogram spectrum from which it is calculated.

This approach is no longer appropriate in our case because a different PSD estimate, namely the Blackman-Tukey estimate, is used. This estimate, as discussed in Section 4.4, has a lower variance and is therefore smoother. Besides just the formant structure is retained hence no harmonics can be found. Therefore the parameters proposed in the original algorithm to select the tonal and non tonal components no longer hold and modifying the decision rules became inevitable. For this reason, the modifications described shortly have been adopted and the resulting masking threshold had no significant differences with the one obtained via the original approach (where a periodogram is used).

To find tonal components, all local maxima are selected. Therefore if $\Phi(k)$ is the Blackman-Tukey PSD estimate obtained via a K-point DFT, then a frequency with index k is added to the tonal maskers list $\mathcal{T}_m$ if

$$\Phi(k) > \Phi(k \pm 1)$$

[4]For example we ruled out using the MPEG model 2 because it requires the phase information to calculate the tonality index. This quantity, by design, is not available in the PSS method.

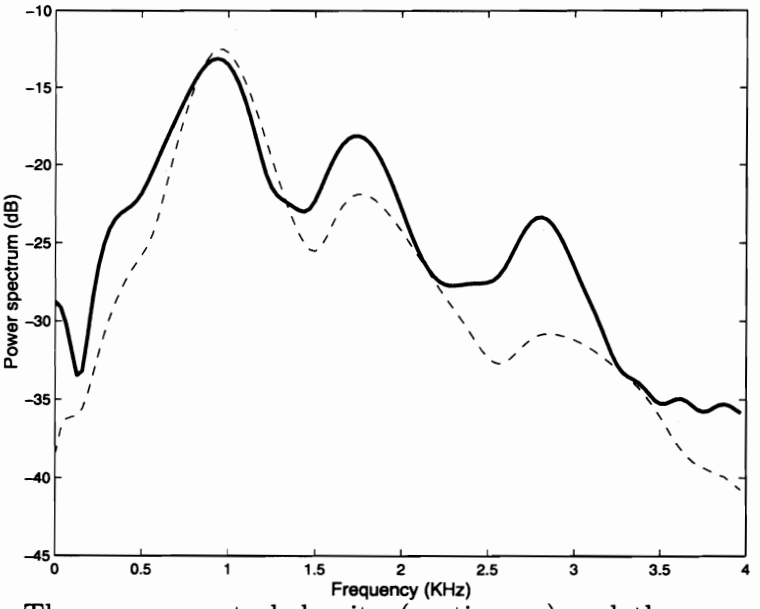


Fig. 5.2 The power spectral density (continuous) and the corresponding masking threshold (dashed) of the vowel /a/.

In the original MPEG model, only frequencies with power 7 dB greater than that of their neighbors are retained in $\mathcal{T}_m$. In our implementation this turned out to be useless due to the smoothness of the spectrum.

After that, all frequencies in the neighborhood of a tonal component are also added to $\mathcal{T}_m$. A neighborhood is chosen to be three spectral lines from any particular tonal component. That is, if k_0 is a tonal component, then so are $k_0 + k$ for $k = -3, \ldots, 3$.

All the remaining frequencies are then labeled as non-tonal and form the non-tonal or noise-like list $\mathcal{N}_m$. To preserve the smoothness of the spectrum, all frequencies are retained for the masking threshold calculation as compared to the one bark resolution of the original algorithm.

Finally the global masking threshold at a particular frequency with index k is calculated in the following way

$$\Phi_{thr}(k) = \sum_{i \in \mathcal{T}_m} T_{tm}(z_i, z_k) + \sum_{i \in \mathcal{N}_m} T_{nm}(z_i, z_k) \quad \text{for } k = 0, \ldots, K/2 \tag{5.11}$$

where $T_{tm}(z_i, z_k)$ and $T_{nm}(z_i, z_k)$ are the individual masking thresholds at z_k barks due to the tonal and non-tonal masking components, respectively, located at z_i barks. The

mapping from linear frequency to bark domain is done using (2.1) as discussed in Chapter 2, that is

$$z_k = z\left(\frac{k}{K}F_s\right) \tag{5.12}$$

where F_s is the sampling rate. Individual tonal and non-tonal masking thresholds are calculated as explained is Section 2.3.2. An example of a masking threshold obtained using the above described algorithm is shown in Figure 5.2 for the vowel /a/.

Note that since this masking threshold model is based on continuous functions, increasing the DFT size of the used spectrum was found to improve the overall speech enhancement performance.

5.5 Calculating the noise energies

Noise estimation is a crucial step for all speech enhancement methods. As discussed in Section 3.5, a noise estimate can be obtained, and updated, during non-speech activity periods, with the help of a VAD.

In the context of SSA, another difficulty is related to handling colored noise since the original approach has been developed for the white noise case, a situation rarely encountered in practice. In this thesis, colored noise is handled in a similar way to the RQSS method described in Section 3.4.5. In the context of PSS this technique is implemented as described next.

Consider the Raleigh Quotient ξ_i associated with noise covariance matrix estimate $\mathbf{R}_w$ and $\mathbf{u}_i$, the i^{th} eigenvector of the clean speech covariance matrix. ξ_i, which is actually the noise energy along the i^{th} spectral direction, is calculated according to (3.91) which is repeated here for convenience

$$\xi_i = \mathbf{u}_i^T \mathbf{R}_w \mathbf{u}_i \tag{5.13}$$

This energy is used in the gain function instead of the constant noise variance σ^2. Using a similar procedure to the one described in Section 4.5, (5.13) can be written in a similar way to (4.4) as follows

$$\xi_i = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_w(\omega)|V_i(\omega)|^2 d\omega \tag{5.14}$$

where $\Phi_w(\omega)$ is the PSD estimate of $\mathbf{w}$. Recall that $\mathbf{R}_w = \text{Toeplitz}(r_w(0), \ldots, r_w(P-1))$ and $\Phi_w(\omega)$ is the DTFT of the noise autocorrelation estimate $r_w(p)$. Following the filterbank

interpretation of the SSA, presented in Section 4.2, ξ_i is actually the noise energy in the i^{th} subband. This interpretation can further justify the use of the Raleigh Quotient to handle colored noise, as discussed earlier.

In matrix notation (5.14) can be expressed as

$$\boldsymbol{\xi} = \frac{1}{K}\mathbf{V}^T\boldsymbol{\Phi}_w \tag{5.15}$$

where $\boldsymbol{\xi} = [\xi_1, \xi_2, \ldots, \xi_Q]^T$ and the matrix $\mathbf{V}$, defined in (4.25) and (4.29), is obtained using a K-point DFT of the eigenvectors of the clean covariance matrix $\mathbf{R}_s$. The vector $\boldsymbol{\Phi}_w = [\Phi_w(0), \ldots, \Phi_w(K-1)]^T$ is chosen to be the Blackman-Tukey PSD estimate of the noise calculated using a K-point DFT. Although (5.15) and (5.13) are not mathematically equivalent, experimental results showed no significant impact of this design decision on the ultimate noise reduction performance.

Computing the noise energies, ξ_i, with (5.15) is preferred to (5.13) because it requires less arithmetic operations. This is because the matrix $\mathbf{V}$ is also needed in the masking threshold computation phase and there will be no additional cost in using it here.

5.6 The overall PSS algorithm

In this section we describe in detail all the steps required to implement the proposed PSS method. Figure 5.3 shows a block diagram of this method. The role of every block is explained next. Recall that, to reduce the computational load, PSS is implemented according to the frame by frame scheme described in Section 5.2. Hence, unless explicitly mentioned, all the steps described next are performed for one length-L frame.

1) Noise Estimation

The role of this block is to provide a noise estimate to be used by PSS. As we have mentioned earlier, during non-speech activity periods, an autocorrelation function and a PSD estimate of the noise, can be obtained. Both the autocorrelation sequence $r_w(p)$ and the PSD estimate $\boldsymbol{\Phi}_w$ should be calculated. The latter is obtained using the Blackman-Tukey estimate by multiplying $r_w(p)$ with a length $2P-1$ Bartlett window.

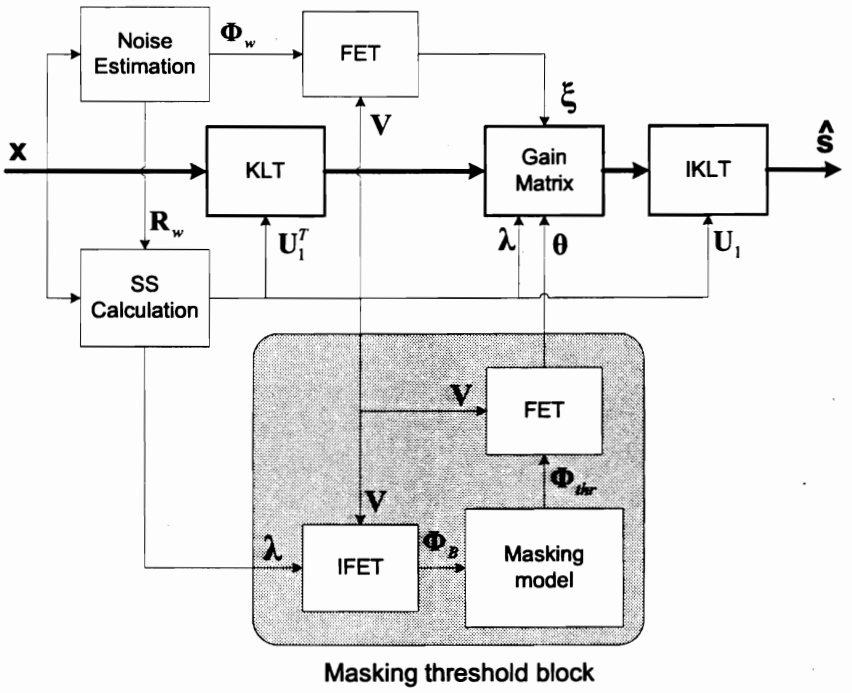


Fig. 5.3 Block diagram of the proposed perceptual signal subspace method

2) Calculating the signal subspace:

According to the discussion of Section 4.3 and our initial experimental tests, the best performance would be achieved using a KLT calculated from a clean speech covariance matrix. Therefore, in PSS, we estimate the clean signal autocorrelation function by subtracting the estimated noise autocorrelation function from the noisy signal autocorrelation function as follows

$$r_s(p) = r_x(p) - r_w(p)$$

The clean signal covariance matrix is obtained as $\mathbf{R}_s = \text{Toeplitz}\{r_s(0), \ldots, r_s(P-1)\}$. Next the EVD of the matrix $\mathbf{R}_s$ is calculated yielding the eigenvalue vector $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_Q]^T$, the eigenvector matrix $\mathbf{U}_1$ and the corresponding matrix $\mathbf{V}$ as discussed in section 4.5. In this case, $\mathbf{R}_s$ is not guaranteed to be positive definite hence the rank Q of $\mathbf{R}_s$ is chosen to be the number of strictly positive eigenvalues of $\mathbf{R}_s$.

3) The masking threshold:

In this step, we use the IFET, equation (4.28), to obtain a clean speech PSD estimate

$$\mathbf{\Phi}_s = \frac{1}{P}\mathbf{V}\boldsymbol{\lambda}$$

This estimate is used to calculate the masking threshold $\mathbf{\Phi}_{thr}$ as described in Sections 2.3.2 and 5.4. The masking energies $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_Q]^T$ are then obtained using the FET (4.30),

$$\boldsymbol{\theta} = \frac{1}{K}\mathbf{V}^T\mathbf{\Phi}_{thr}$$

As discussed earlier a DFT with size $K = L$ has been used in order to obtain masking energies resulting in a better enhancement performance.

4) The KLT:

For every vector within the current frame, the signal coefficients in the signal subspace are obtained by multiplying the input vector $\mathbf{x}$ by the KLT matrix $\mathbf{U}_1^T$.

5) The gain matrix:

The so obtained signal coefficients are multiplied by a diagonal gain matrix $\mathbf{G}$ to reduce the undesired interfering noise along every eigen direction. The gain matrix entries are calculated as follows.

Using FET, obtain the noise energies $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_Q]^T$ as follows

$$\boldsymbol{\xi} = \frac{1}{K}\mathbf{V}^T\mathbf{\Phi}_w$$

where the noise estimate $\mathbf{\Phi}_w$ is obtained from the noise estimation module. The gain function is then calculated as

$$g_i = e^{-\nu\xi_i/\min(\lambda_i,\theta_i)} \quad \text{for } i = 1, \ldots, Q$$

and the gain matrix is given by $\mathbf{G} = \text{diag}(g_1, \ldots, g_Q)$.

6) The IKLT:

The enhanced signal vector is finally recovered in the signal subspace using the inverse KLT matrix $\mathbf{U}_1$. Specifically, the clean speech vector estimate is given by

$$\hat{\mathbf{s}} = \mathbf{U}_1 \left[\mathbf{G}\mathbf{U}_1^T\mathbf{x}\right]$$

Every output vector is then multiplied by a Hanning window and using the overlap-add synthesis technique, one frame of speech is recovered. Multiplying this frame by a second (larger) Hanning window and again using the overlap-add synthesis technique, the overall clean speech estimate $\hat{s}(n)$ is obtained.

5.7 Summary

In this chapter, using the Frequency to Eigendomain Transformation (FET), we presented a new perceptual signal subspace method designed according to the masking properties of the human auditory system. This method modifies the gain function so that the noise be suppressed according to the ratio of its energy to the masking energy along a particular eigen direction. In addition to white noise, the proposed method is also capable of handling the more general and more practical case of colored noise.

To enhance the method, we also proposed a novel frame-based technique for EVD calculation which takes into account the stationarity assumption of the speech signal for a long enough period of time. This is important since the major drawback of SSA based methods is currently their relatively high computational load.

In Chapter 8, we provide some experimental results which test the performance of the proposed PSS method and reveal its superiority over competing methods. The computational savings arising from the FBEVD technique will also be assessed.

Chapter 6

The Multi-Microphone Approach

Single channel speech enhancement methods have usually been the most appealing approach in practice. Their popularity stems mainly from their low cost and ease of implementation. However, their performance still does not meet the expectations of the ever demanding speech industry market. Indeed, the need for more satisfactory speech quality and intelligibility under, for example, very harsh acoustic conditions such as in-car hands free applications, has steered the researchers' attention to multi-microphone techniques where the added speech acquisition channels seem to offer better solutions to the speech enhancement problem.

Nowadays, microphone arrays find applications in different areas including teleconferencing [21], hands-free telephony [54, 117], hearing aids [93] and speech recognition [4, 98]. They were successfully used to reduce both noise and reverberation [1, 3, 47, 111, 117, 137]. Promising results in a combined cancellation of noise and echo have also been reported [26, 36].

Depending on the underlying noise field, whether diffuse or directional, several microphone array methods have been proposed in the literature. These methods usually exploit both the spatial and temporal redundancy in the acquired speech signals to filter out the interfering noises. The use of the signal subspace approach in microphone arrays, however, has not received much attention. This is in contrast to other signal processing areas, such as array processing, where the signal subspace tool is commonly used. For instance, the MUSIC algorithm, which is a signal subspace based technique, has been considered as a breakthrough in direction of arrival (DOA) estimation research [134, 135].

In this chapter, we extend the single channel signal subspace technique for speech enhancement into a multi-microphone design. The proposed method, called the Multi-microphone Eigen-Domain Averaging method (MEDA), exploits a property of the covariance matrix of the signals gathered from the different available microphones to improve the noise suppression capabilities of the SSA. Indeed, the structure of the covariance matrix is taken advantage of to accomplish averaging in the eigen domain resulting in filter coefficients less vulnerable to the environmental conditions.

By design, the MEDA is mainly intended for diffuse noise field applications where its performance has been experimentally found to be considerably higher than competing methods designed for such noise fields. As compared to this category of methods, the MEDA significantly reduces the residual noise level while maintaining a similar signal distortion. Under directional noise, the performance of MEDA is again superior to these methods although it is not as good as specialized methods.

This chapter starts by presenting the problem of multi-microphone speech enhancement and then a description of the most common noise field models available in the literature is given. Next, a survey of some popular multi-microphone methods is provided. After that, the novel MEDA method is described and analyzed. Experimental results assessing the performance of MEDA and comparing it with some competing methods are given in Section 8.5.

6.1 Problem formulation

Consider a linear array of M microphones where the distance between every pair with indices m and l is given by d_{ml}. The microphones are assumed to be omni-directional and having a flat frequency response equal to one[1].

Besides, we assume a far-field situation where plane wave propagation can be considered to be valid. In this case the signal attenuation can be assumed to be equal for all microphones. In a far-field situation, the sound sources should be far enough from the microphone array and usually the following condition should be satisfied [115]

$$r > \frac{F_s d_{M1}^2}{c} \tag{6.1}$$

[1]This ideal situation is however rarely met in practice and the microphone response can be modeled as a convolutional noise.

where r is the distance from the source to the center of the array, F_s is the sampling rate and c is the speed of sound. The distance d_{M1} between the 1^{st} and the M^{th} microphone represents the total size of the array. Violating this conditions implies a near-field situation in which spherical wave propagation and signal attenuation should be considered.

In general the direct path signal will make an angle θ with the array axis. An angle of $\theta = 0°$ is called broadside whereas an angle of $\theta = 90°$ or $\theta = -90°$ is called end-fire. The angle θ is called the direction of arrival (DOA) of the sound source. A DOA different from zero degrees (i.e. non broadside) will result in the signal arriving at two different microphones at different times. The time delay between two microphones with indices m and l is given by

$$\tau_{ml} = \frac{d_{ml} \sin\theta}{c} \tag{6.2}$$

Generally, the time delay is measured with respect to one particular microphone, say having index $m = 1$, and τ_m then represents the time delay between microphone 1 and microphone m. Hence, we have $\tau_{ml} = \tau_m - \tau_l$ and obviously $\tau_1 = 0$.

6.2 Time delay compensation

Time delay estimation is very crucial for the performance of microphone arrays as it is needed to steer the microphone array towards one specific direction (the look direction) in order to synchronize the desired speech signal from the direct path over all microphones. This is usually done by compensating the incurred time delays between the different channels by passing the acquired microphone signals through a time delay compensation module.

Let $y_m(n)$ be the sampled discrete speech signal at the output of the m^{th} microphone. The signal at the output of the time delay compensation module is then given by

$$x_m(n) = y_m(n - \delta_m) \quad \text{for } m = 1, \ldots, M \tag{6.3}$$

where $\delta_m = -\tau_m F_s$. Since the time delays are usually non integers, the above operation is generally implemented using interpolation filters [86].

Doing so, $x_m(n)$ can then be written as

$$x_m(n) = s(n) + w_m(n), \quad m = 1, \ldots, M \tag{6.4}$$

where $s(n)$ is the desired speech signal and $w_m(n)$ consists of all interfering signals including additive noise and possibly convolutional noise due to reverberation and a non flat microphone frequency response. In the forthcoming sections however, the convolutional noise is ignored and the noise term is assumed to be uncorrelated with the speech signal.

Due to its importance, several methods for *time delay estimation* (TDE) or DOA estimation have been proposed in the literature. One of the most popular methods employed is the generalized cross-correlation method (GCC) [100]. The GCC owes its popularity to its simplicity and robustness. One of the GCC based estimators, the PHAse Transform estimator (PHAT), has been recently found to be optimal under reverberant conditions [58]. Actually, reverberation is found to limit the performance of the GCC [18]. To cope with this problem, several other methods have been proposed, for example [14, 13, 122, 146]. A good survey on time delay estimation can be found in [16].

A survey of other DOA estimation methods, which are not necessarily designed for speech signals, can also be found in [101]. Among these methods are the high resolution subspace based methods which involve the EVD of the spatial covariance matrix. These methods include the MUSIC [134, 135] and ESPRIT [133] algorithms. These methods, however, are mainly designed for narrowband signals and thus are not appropriate for speech signals. Further processing is therefore suggested such as the use of Coherent Signal Subspace (CSS) methods [161]. In this approach, focusing matrices are used to align narrow-band components within the receiver bandwidth prior to forming covariance matrix estimates at each frequency [153]. Another approach for wideband direction of arrival estimation are the frequency independent beamforming methods [162, 106].

In what follows, it will be assumed that "perfect" time delay compensation pre-processing has been performed so that the desired signal can be assumed to be exactly synchronized over all microphones. Only the signal $x_m(n)$ will then be considered and, for simplicity, will be referred to as the output of the m^{th} microphone. Equivalently, this implies that it is assured that the direct path desired signal is impinging on the array from broadside ($\theta = 0°$).

6.3 Noise field models

Consider an array consisting of M microphones where the signal at the output of the m^{th} microphone is given by $x_m(n)$ for $m = 1, \ldots, M$. The noise field is usually characterized

by the so-called spatial coherence function defined as [117]

$$\Gamma_{ml}(\omega) = \frac{\Phi_{ml}(\omega)}{\sqrt{\Phi_m(\omega)\Phi_l(\omega)}} \tag{6.5}$$

where $\Phi_{ml}(\omega)$ is the cross-power spectrum of the signals at microphone m and microphone l and $\Phi_m(\omega)$ and $\Phi_l(\omega)$ are their respective PSD's. Often, just the magnitude squared coherence (MSC), given by $C_{ml}(\omega) = |\Gamma_{ml}(\omega)|^2$, is used. Note that $0 \leq |\Gamma_{ml}(\omega)|^2 \leq 1$.

Of interest, also, is the average coherence function of the noise field for all sensor pairs $m \neq l$, given by

$$\Gamma(\omega) = \frac{2}{M(M-1)} \sum_{m=1}^{M-1} \sum_{l=l+1}^{M} |\Gamma_{ml}(\omega)| \tag{6.6}$$

where the magnitude operator is used to ensure that the average coherence function is a real positive quantity[2]. This function measures the amount of coherence (or correlation) between the signals at two microphones at a particular frequency. A coherence of zero indicates that the two signals are completely uncorrelated whereas a value of one indicates a total correlation.

6.3.1 Incoherent noise field

As indicated from its name, an incoherent field occurs when the signals at two microphones are completely uncorrelated at all frequencies. Hence the coherence function would have a constant value equal to zero. This may occur for example with the internal noise of the microphones.

For most noise fields in practice, however, this is rarely satisfied for all frequencies. While an incoherent noise field assumption is often used for microphone array filter design, the best scenario that can occur is the so called diffuse noise field, described in the next section.

6.3.2 Diffuse noise field

A diffuse noise field (or ambient noise) occurs when there is a superposition of an infinite number of plane waves, due for example to an infinite number of sound sources, impinging

[2]Sometimes the real part can also be used instead.

on the array from different directions. This has been found to be a suitable model in reverberant enclosures where the large number of wall reflected signals would eventually form a diffuse noise field [103, 58]. This model is also found to be valid to characterize the noise field inside cars [52].

In a diffuse noise field the coherence is real valued and is given by [117]

$$\begin{cases} Re\{\Gamma_{ml}(\omega)\} & = \text{sinc}\left(\frac{\omega d_{ml}}{c}\right) \\ Im\{\Gamma_{ml}(\omega)\} & = 0 \end{cases} \tag{6.7}$$

where d_{ml} is the distance between the m^{th} and l^{th} microphones and c is the speed of sound. In this thesis, the sinc function is defined as $sinc(x) \triangleq sinx/x$.

Therefore, two microphone signals can be considered to be uncorrelated if they are only composed of high frequencies. At low frequencies, on the other hand, the coherence increases approaching unity as the frequency gets closer to 0 Hz. Indeed, to obtain a low coherence, the inter-microphone distance d_{ml} should be set to meet the condition

$$d_{ml} > \frac{c}{2f_{\min}} \tag{6.8}$$

where $f_{\min}$ is the lowest frequency of interest.

Increasing the distance however, may result in violation of the far-field assumption and in making the array more vulnerable to steering errors.

6.3.3 Coherent noise field

The third noise filed type which is commonly encountered in practice is the coherent noise field. A coherent noise field occurs when there is one sound source impinging from one direction at an angle θ. This situation is also often referred to as a directional sound source. In this case the coherence function is given by [117],

$$\begin{cases} Re\{\Gamma_{ml}(\omega)\} = & \cos\left(\frac{\omega\cos(\theta)d_{ml}}{c}\right) \\ Im\{\Gamma_{ml}(\omega)\} = & -\sin\left(\frac{\omega\cos(\theta)d_{ml}}{c}\right) \end{cases} \tag{6.9}$$

Note that the magnitude of this coherence is equal to one for all frequencies.

6.4 Multi-microphone methods

In this section we describe some of the most popular multi-microphone methods of speech enhancement. The detailed description of these techniques can be found in the references cited herein.

6.4.1 Fixed beamforming

In the conventional delay-and-sum (DS) beamformer [44, 45, 95, 96], the delayed microphone signals at the output of the delay compensation module are weighted and summed as follows

$$z(n) = \frac{1}{M} \sum_{m=1}^{M} a_m x_m(n) \tag{6.10}$$

where a_m's are some fixed (non-adaptive) weights used to give the beam pattern a specific desired shape. In a near-field situation these weights can be used to compensate for the signal attenuation and to equalize the level of the direct path signal over all channels. They can also be used as a remedy to microphones with different gains. In this thesis however, this weighting, also often referred to as shading, is not considered and is supposed to be unnecessary.

The main lobe width of the beam pattern is frequency dependent and becomes wider at low frequencies for a given fixed inter-microphone distance. This results in poor performance at low frequencies. This phenomenon lead to the development of frequency invariant beamformers [51, 20]. Frequency invariant beamforming can also be achieved using nested arrays [45, 95]. In such microphone arrays, a subarray is used for every frequency band so that the frequency and inter-microphone distance product remains relatively constant. Clearly, sub-arrays with larger inter-microphone distances are used for low frequencies and sub-arrays with smaller distances, are used for high frequencies. Unfortunately, a large number of microphones is then needed to achieve an acceptable spatial selectivity which is not practical in general in terms of spatial placement and the total cost of the whole system.

The noise reduction factor of the DS beamformer is given by [120]

$$NRF(\omega) = \frac{1}{\frac{1}{M} + \left(1 - \frac{1}{M}\right) \Gamma(\omega)} \tag{6.11}$$

where M is the number of microphones and $\Gamma(\omega)$ is the average coherence function given in (6.6). Note that the noise reduction factor is defined as the ratio of the noise energy at the output of the beamformer to the energy of the original corrupting noise at the input of one microphone[3]. It can be seen that the performance of the DS beamformer is proportional to the number of microphones. For example for an incoherent noise field ($\Gamma(\omega) = 0$), the noise reduction factor is $NRF(\omega) = M$. This fact confirms that large microphone arrays are required to achieve an acceptable performance.

Unfortunately, in most applications of interest, just a few microphones can actually be used. For example in car applications[4], the car makers may refuse to install more than two microphones. The reason for that is mainly the spatial constraints which make it complicated to employ a larger number of microphones. In addition to that, the total cost of the whole system can increase in a way to make the achieved performance improvement too expensive to be desired. For this reason, fixed beamformers are usually used in conjunction with other techniques to maintain a satisfactory performance despite the small number of microphones used. Such techniques, discussed next, are mainly adaptive beamforming and adaptive postfiltering.

6.4.2 Adaptive beamforming

To overcome the limitations of the DS beamformer in speech enhancement applications, adaptive beamforming has been proposed [92, 91, 141]. These systems are usually based on the so called Generalized Sidelobe Cancelor (GSC) [55], which results from the transformation of a constrained optimization problem into a non-constrained one [154]. As illustrated in Figure 6.1, the GSC consists of a conventional fixed DS beamformer and a blocking matrix. The blocking matrix is designed to block the desired signal in the look direction in order to obtain a speech free estimate of the noise. The output of the blocking matrix is used to adaptively steer a null in the directions of the interference. The GSC, however, suffers from what is known as the signal cancellation effects. This takes place when there is leakage of the desired signal into the blocking matrix resulting in serious signal distortion [163]. For example in reverberant rooms the GSC is known to have considerable performance limitations [8].

[3]The noise energy is commonly assumed to be equal at all microphones.

[4]such as hands-free telephony which is increasingly being enforced by law in many countries for security reasons.

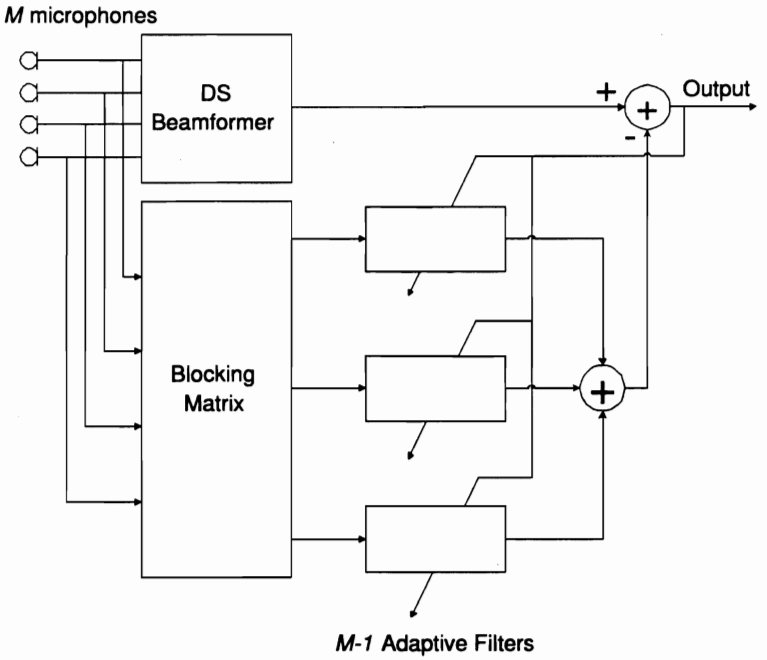


Fig. 6.1 The Griffiths-Jim beamformer or the Generalized Side-lobe Cancellor (GSC).

A classical noise reduction method which can also be viewed as a form of adaptive beamforming is the *adaptive noise canceling* (ANC) method. ANC uses a dual-microphone concept which can be successfully used for speech enhancement if there is a high coherence of the noise in the two channels while the speech is present in just one of these channels [30, 52, 158, 159, 164]. However, in most practical situations, both requirements can not be met at the same time.

While adaptive beamforming shows in general good performance for directional interference (coherent noise field) it is no better than a DS beamformer in diffuse or incoherent fields [8]. For example, in a car environment, the noise field (mainly due to the engine) is considered to be diffuse. Therefore, adaptive beamforming is not expected to have a good noise reduction performance. This observation is also valid for ANC [52].

6.4.3 Adaptive postfiltering

For diffuse noise fields, adaptive postfiltering has been proposed as a better alternative for noise reduction. Actually, these systems usually assume an incoherent noise field to derive

the transfer function of the suppression filter. This assumption is reasonable for high frequencies but is untrue for low frequencies as discussed in Section 6.3.2. For this reason, the noise reduction capabilities at low frequencies is usually no better than a conventional delay-and-sum beamformer. However, adaptive postfiltering remains the best design to use in diffuse fields and usually further processing is applied to enhance performance at low frequencies.

Depending on the design criterion, the post-filter can take different forms: a Wiener filter [117, 170], an LMS-type adaptive filter [169], a coherence function based filter [105, 104], a combination of coherence and Wiener filtering [114] and a combination of Wiener filtering and spectral subtraction [120]. A block diagram of this method is shown in Figure 6.2.

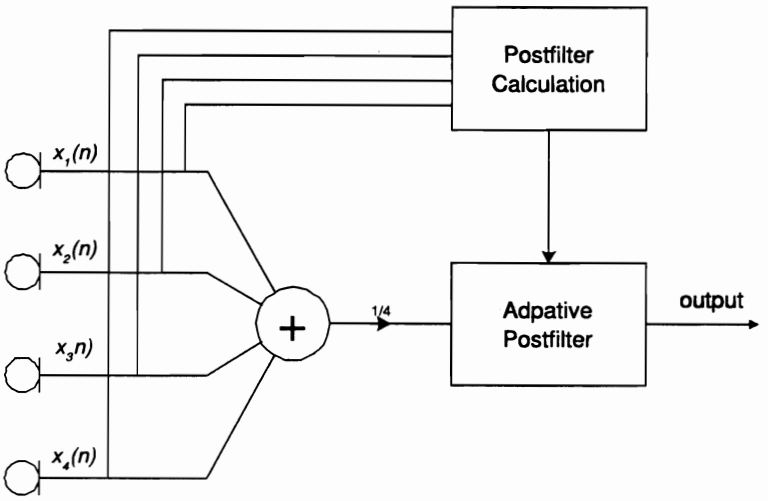


Fig. 6.2 A block diagram of an array of four microphones with an adaptive postfilter.

The adaptive postfiltering technique consists first of a conventional DS beamformer[5]

$$z(n) = \frac{1}{M}\sum_{m=1}^{M} x_m(n) = s(n) + \frac{1}{M}\sum_{m=1}^{M} w_m(n) \tag{6.12}$$

Then, to further reduce the remaining noise components at the beamformer output, $z(n)$ is post-filtered via a Wiener filter having the following frequency response

$$H(\omega) = \frac{\Phi_{zs}(\omega)}{\Phi_z(\omega)} \tag{6.13}$$

[5]Weighting as in (6.10) can still be used to improve the directivity of the array [117].

where $\Phi_{zs}(\omega)$ is the cross-power spectrum of $s(n)$ and $z(n)$, and $\Phi_z(\omega)$ is the power spectrum of $z(n)$. The noise and the speech are assumed to be uncorrelated, so using (6.12) we have $\Phi_{zs}(\omega) = \Phi_s(\omega)$ and $H(\omega)$ can then be written as

$$H(\omega) = \frac{\Phi_s(\omega)}{\Phi_z(\omega)} \tag{6.14}$$

In practice, however, $\Phi_s(\omega)$ is not available and needs to be estimated. This estimate is obtained by exploiting the signals from the different available channels in the following way

$$\hat{\Phi}_s(\omega) = \frac{2}{M(M-1)} Re\left\{ \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \Phi_{x_i x_j}(\omega) \right\} \tag{6.15}$$

where the real part is used in order to reduce any possible coherent noise components in the input channels[6] [170]. This estimate is used to replace $\Phi_s(\omega)$ in (6.14). In this formulation, it is assumed that an incoherent noise field (which is usually employed as an approximation to a diffuse field) is used.

Alternatively, the postfilter can be estimated using the average coherence function, defined in (6.6), over all microphone pairs. The frequency response of the filter will then take the following form [104],

$$H(\omega) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \frac{|\Phi_{x_i x_j}(\omega)|}{\sqrt{\Phi_{x_i}(\omega)\Phi_{x_j}(\omega)}} \tag{6.16}$$

Due to the underlying noise field assumption, this technique is suitable in applications where the environment can be modeled by a diffuse noise field, such as in cars or reverberant rooms.

6.4.4 The Multi-microphone SVD method

So far, just a few attempts to use the signal subspace approach in a microphone array for speech enhancement have been reported. In [4] the EVD of the spatial covariance matrix is used to design a minimum variance (MV) beamformer to improve the performance of speech recognition systems in noisy environments. In this approach the so-called coherent

[6] Comparable results can also be obtained by taking the magnitude instead of the real part [117].

signal subspace approach using a focusing matrix [161] is employed.

A second method, which is more relevant to the context of this thesis as will be seen shortly, is the so called Multi-microphone Singular Value Decomposition (MSVD) method [35, 34]. We next provide a brief description of the MSVD.

Consider the PM-dimensional noisy observation vector $\bar{\mathbf{x}}(n)$ formed by stacking the different microphone signal vectors above each other as follows

$$\bar{\mathbf{x}}(n) = [\mathbf{x}_1(n)^T, \ldots, \mathbf{x}_M^T(n)]^T = \bar{\mathbf{s}}(n) + \bar{\mathbf{w}}(n) \tag{6.17}$$

where $\bar{\mathbf{s}}(n)$ is the clean speech vector and $\bar{\mathbf{w}}(n)$ is the noise vector. M is the number of microphones and P is the dimension of the sub-vectors given by

$$\mathbf{x}_m(n) = [x_m(n), x_m(n-1), \ldots, x_m(n-P+1)]^T$$

From these stacked vectors, a $L \times MP$ data matrix $\bar{\mathbf{X}}(n)$ at time index n is formed in the following way,

$$\bar{\mathbf{X}}(n) = \begin{bmatrix} \bar{\mathbf{x}}^T(n) \\ \bar{\mathbf{x}}^T(n+1) \\ \vdots \\ \bar{\mathbf{x}}^T(n+L-1) \end{bmatrix} \tag{6.18}$$

or equivalently

$$\bar{\mathbf{X}}(n) = [\mathbf{X}_1(n) \;\; \mathbf{X}_2(n) \;\; \ldots \;\; \mathbf{X}_M(n)] \tag{6.19}$$

where $\mathbf{X}_m(n)$, for $m = 1, \ldots, M$, is a $L \times P$ data matrix obtained from $L + P - 1$ data samples as follows,

$$\mathbf{X}_m(n) = \begin{bmatrix} x_m(n) & x_i(n-1) & \ldots & x_m(n-P+1) \\ x_m(n+1) & x_m(n) & \ldots & x_m(n-P+2) \\ \vdots & \vdots & & \vdots \\ x_m(n+L-1) & x_m(n+L-2) & \ldots & x_m(n+L-P) \end{bmatrix} \tag{6.20}$$

that is,

$$\mathbf{X}_m(n) = \begin{bmatrix} \mathbf{x}_m^T(n) \\ \mathbf{x}_m^T(n+1) \\ \vdots \\ \mathbf{x}_m^T(n+L-1) \end{bmatrix} \tag{6.21}$$

Hereafter, the time index n will be dropped for simplicity.

In a similar way, a $L_w \times MP$ noise data matrix $\bar{\mathbf{W}}$ is formed using L_w, not necessarily consecutive, stacked vectors gathered during non speech activity periods.

Consider now the Generalized Singular Value Decomposition (GSVD) of $\bar{\mathbf{X}}$ and $\bar{\mathbf{W}}$ given by

$$\bar{\mathbf{X}} = \bar{\mathbf{V}}_x \text{diag}\{\sigma_i\} \bar{\mathbf{U}}^T \tag{6.22}$$

$$\bar{\mathbf{W}} = \bar{\mathbf{V}}_w \text{diag}\{\eta_i\} \bar{\mathbf{U}}^T \tag{6.23}$$

where $\bar{\mathbf{V}}_x$ and $\bar{\mathbf{V}}_w$ are two matrices with orthonormal columns and $\bar{\mathbf{U}}$ is an invertible, but not necessarily orthogonal, matrix. Accordingly, an optimal Wiener filter minimizing the residual error signal energy can be obtained as [35]

$$\bar{\mathbf{H}} = \bar{\mathbf{U}}^{-T} \text{diag}\{1 - \frac{L}{L_w} \cdot \frac{\eta_i^2}{\sigma_i^2}\} \bar{\mathbf{U}}^T \tag{6.24}$$

The estimate of the clean speech data matrix is then obtained as $\hat{\bar{\mathbf{S}}} = \bar{\mathbf{X}}\bar{\mathbf{H}}$.

In the (deterministic) signal subspace methods which use the SVD of a data matrix such as in [61, 84], averaging along the sub-diagonals of the resulting clean speech data matrix estimate is carried out in order to obtain the final enhanced speech signal. However, in [35], it is suggested that this averaging step is not optimal. The optimal filter is rather obtained by selecting the column of $\bar{\mathbf{H}}$ in (6.24) which corresponds to the smallest element on the diagonal of the matrix $\bar{\mathbf{W}}^T\bar{\mathbf{W}}\bar{\mathbf{H}}$. Since finding such an element is costly in terms of the required computations, it is claimed that picking the middle column is enough to obtain a satisfactory result.

MSVD is found to outperform other beamforming methods, namely fixed beamforming and the GSC. This performance superiority is due to the fact that the implemented signal subspace filter can be viewed as a cascade of a spatial filter (beamformer) and a Wiener-like

postfilter which depends on the SNR at the output of the beamformer. The beamformer stage automatically steers the array in the direction of the desired speech signal and places a null in the direction of the interference. Unlike the GSC, for instance, the gain of the main lobe is not equal to one and will generally depend on the signal to noise ratio giving rise to a postfilter like behavior. This property leads to the reported improved performance of the MSVD over the GSC.

Another merit of the MSVD method is that it makes no assumptions on the DOA of the desired speech signal making it less vulnerable to steering errors. However, this is achieved at the cost of a degradation in performance as the number of interfering signals increases [34]. Consequently, this drawback resulted in a poor performance under reverberant conditions with the degradation becoming more serious as the reverberation time increases. Indeed, this behavior was confirmed in our experimental results as will be seen in Section 8.5. Therefore, it can be concluded that while MSVD remains a robust speech enhancement tool under directional noise, its capabilities remain rather modest under diffuse noise fields.

6.5 The multi-microphone EVD approach

In this section we present an extension of the single microphone SSA into a multi-microphone design. The method developed here, although by itself can be considered as a novelty of this thesis, mainly serves as a basis for the new MEDA method which will be presented in the Section 6.6.

Consider a microphone array consisting of M microphones. Define, as in (6.17), the PM-dimensional composite input vector $\bar{\mathbf{x}}$,

$$\bar{\mathbf{x}} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_M^T]^T = \bar{\mathbf{s}} + \bar{\mathbf{w}} \tag{6.25}$$

where $\bar{\mathbf{s}}$ is the clean speech vector and $\bar{\mathbf{w}}$ is the noise vector defined in a similar way.

Assume that the noise and speech are uncorrelated so that the noisy Composite Covariance Matrix (CCM), $\bar{\mathbf{R}}_x = E\{\bar{\mathbf{x}}\bar{\mathbf{x}}^T\}$, of the noisy composite vector can be written

as

$$\bar{\mathbf{R}}_x = \begin{bmatrix} \mathbf{R}_{x,11} & \mathbf{R}_{x,12} & \cdots & \mathbf{R}_{x,1M} \\ \mathbf{R}_{x,21} & \mathbf{R}_{x,22} & \cdots & \mathbf{R}_{x,2M} \\ \vdots & & \ddots & \vdots \\ \mathbf{R}_{x,M1} & \cdots & \cdots & \mathbf{R}_{x,MM} \end{bmatrix} = \bar{\mathbf{R}}_s + \bar{\mathbf{R}}_w \tag{6.26}$$

where $\bar{\mathbf{R}}_s$ and $\bar{\mathbf{R}}_w$ are the clean speech and noise CCM's respectively. $\mathbf{R}_{x,ml}$ is the cross-covariance matrix between microphones m and l, i.e. $\mathbf{R}_{x,ml} = E\{\mathbf{x}_m \mathbf{x}_l^T\}$.

A far-field situation is assumed so that signal attenuation is constant over all microphones. Besides recall, as mentioned in Section 6.2, that we assume perfect time delay compensation has already been applied so that the desired direct path speech signal is in phase over all M microphones. Therefore, similar to (6.4), the noisy observation vector can be written as

$$\mathbf{x}_m = \mathbf{s} + \mathbf{w}_m, \quad \text{for } m = 1, \ldots, M \tag{6.27}$$

which implies that the clean speech CCM can be written as

$$\bar{\mathbf{R}}_s = \begin{bmatrix} \mathbf{R}_s & \cdots & \mathbf{R}_s \\ \vdots & \ddots & \vdots \\ \mathbf{R}_s & \cdots & \mathbf{R}_s \end{bmatrix} \tag{6.28}$$

where $\mathbf{R}_s = E\{\mathbf{s}\mathbf{s}^T\}$ is the covariance matrix of the clean speech vector $\mathbf{s}$.

At this point we introduce the $MP \times P$ matrix $\mathbf{C} = [\mathbf{I}_P, \ldots, \mathbf{I}_P]^T$ where $\mathbf{I}_P$ is a $P \times P$ identity matrix. This matrix will be frequently used in this chapter to simplify the notation and to facilitate the forthcoming derivations. The properties of this matrix are discussed in detail in appendix A.

The matrix $\bar{\mathbf{R}}_s$ can then be written using $\mathbf{C}$ as follows

$$\bar{\mathbf{R}}_s = \mathbf{C}\mathbf{R}_s\mathbf{C}^T \tag{6.29}$$

It can be readily seen that if $\mathbf{R}_s$ has rank $Q \leq P$ then $\bar{\mathbf{R}}_s$ will also have rank Q. Thus, if $\mathbf{R}_s = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}}_s\bar{\mathbf{V}}^T$ is the EVD of $\bar{\mathbf{R}}_s$ then there will be just Q non-zero eigenvalues and we can write $\bar{\mathbf{V}} = [\bar{\mathbf{V}}_1 \;\; \bar{\mathbf{V}}_2]$ where $\bar{\mathbf{V}}_1$ is a $MP \times Q$ matrix spanning the signal subspace and having as columns the eigenvectors of $\bar{\mathbf{R}}_s$ corresponding to the non-zero eigenvalues[7].

[7]Recall that the eigenvalues are as usual sorted in a decreasing order.

6.5.1 Filter design

The objective now is to design a linear filter $\bar{\mathbf{H}}$ in order to estimate the composite clean speech vector $\bar{\mathbf{s}}$ as

$$\hat{\bar{\mathbf{s}}} = [\hat{\mathbf{s}}_1^T, \ldots, \hat{\mathbf{s}}_M^T]^T = \bar{\mathbf{H}}\bar{\mathbf{x}}. \tag{6.30}$$

Similar to the single microphone case, the filter $\bar{\mathbf{H}}$ may be obtained for example as the solution to the following optimization problem in which the residual error energy is minimized,

$$\min_{\bar{\mathbf{H}}} E\{||\bar{\mathbf{r}}||^2\} \tag{6.31}$$

where the residual error signal $\bar{\mathbf{r}}$ is defined as

$$\bar{\mathbf{r}} = \hat{\bar{\mathbf{s}}} - \bar{\mathbf{s}} = \bar{\mathbf{H}}\bar{\mathbf{x}} - \bar{\mathbf{s}} \tag{6.32}$$

The solution to this problem is the classical Wiener filter

$$\bar{\mathbf{H}} = \bar{\mathbf{R}}_s(\bar{\mathbf{R}}_s + \bar{\mathbf{R}}_w)^{-1}. \tag{6.33}$$

Using the EVD of $\bar{\mathbf{R}}_s$, (6.33) becomes

$$\bar{\mathbf{H}} = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}}_s(\bar{\mathbf{\Lambda}}_s + \bar{\mathbf{V}}^T\bar{\mathbf{R}}_w\bar{\mathbf{V}})^{-1}\bar{\mathbf{V}}^T \tag{6.34}$$

Assuming, for now, a spatio-temporal white noise situation in which the noise CCM can be written as $\bar{\mathbf{R}}_w = \sigma^2\mathbf{I}_{MP}$ where σ^2 is the noise variance in each of the M channels. In this case the filter $\bar{\mathbf{H}}$ in (6.34) can be written as

$$\bar{\mathbf{H}} = \bar{\mathbf{V}}_1\bar{\mathbf{G}}\bar{\mathbf{V}}_1^T \tag{6.35}$$

where in general, as in the single microphone case, $\bar{\mathbf{G}}$ is a $Q \times Q$ diagonal gain matrix with entries depending on the signal to noise ratio in every spectral direction. For example the gain coefficients can take the following form,

$$\bar{g}_i = \frac{\bar{\lambda}_{s,i}}{\bar{\lambda}_{s,i} + \mu\sigma^2} \quad \text{for } i = 1, \ldots, Q. \tag{6.36}$$

where $\bar{\lambda}_{s,i}$ is the i^{th} eigenvalue of $\bar{\mathbf{R}}_s$. The control parameter μ is added in order to provide

a tradeoff between the signal distortion and the residual noise level. In this chapter, this gain function is used because it resembles the gain function of the MSVD method which will be used for evaluation.

Based on our previous assumptions of synchronized microphone signals and constant signal attenuation, the clean speech vector estimate is finally obtained by taking the average of the individual sub-vectors,

$$\begin{aligned} \hat{\mathbf{s}} &= \frac{1}{M}\sum_{m=1}^{M}\hat{\mathbf{s}}_m \\ &= \frac{1}{M}\mathbf{C}^T[\bar{\mathbf{H}}\bar{\mathbf{x}}] \end{aligned} \tag{6.37}$$

6.5.2 The spatio-temporal colored noise case

If the noise was spatio-temporally colored, then the noise CCM would not have a diagonal form. To handle such a scenario, only the noise energy in the direction of every eigenvector is accounted for in the filter $\bar{\mathbf{H}}$. This energy can be calculated as

$$\bar{\xi}_i = \bar{\mathbf{v}}_i^T\bar{\mathbf{R}}_w\bar{\mathbf{v}}_i \quad i = 1,\ldots,Q \tag{6.38}$$

which is the Raleigh Quotient associated with the noise CCM $\bar{\mathbf{R}}_w$ and $\bar{\mathbf{v}}_i$, the i^{th} eigenvector of $\bar{\mathbf{R}}_s$. The gain coefficients in (6.36) can then be altered in the following way

$$\bar{g}_i = \frac{\bar{\lambda}_{s,i}}{\bar{\lambda}_{s,i} + \mu\bar{\xi}_i} \quad \text{for } i = 1,\ldots,Q. \tag{6.39}$$

This approach is similar to the one used in the modified SSA method, described in Section 3.4.5 and referred to as the RQSS method. Variants of the RQSS have been reported to handle colored noises better than prewhitening [130, 121]. For this reason the method described in this section will be hereafter referred to as the Multi-microphone RQSS (MRQSS) method.

The validity of the approximation

$$\bar{\mathbf{V}}^T\bar{\mathbf{R}}_w\bar{\mathbf{V}} = \text{diag}\{\bar{\xi}_1,\ldots,\bar{\xi}_M\}, \tag{6.40}$$

used above, will be further addressed in Section 6.6.

6.5.3 Estimating the Covariance matrix

Since the clean speech CCM is not known it needs to be estimated. As in the RQSS method this approximation is given by $\bar{\mathbf{R}}_x - \bar{\mathbf{R}}_w$, where the noise CCM $\bar{\mathbf{R}}_w$ is obtained during non-speech activity periods. The noisy CCM $\bar{\mathbf{R}}_x$ can be estimated in two different ways as explained below.

In the first approach, the covariance matrix is calculated as[8]

$$\bar{\mathbf{R}}_x = \frac{1}{L}\bar{\mathbf{X}}^T(n)\bar{\mathbf{X}}(n) \tag{6.41}$$

where $\bar{\mathbf{X}}^T(n)$ is a data matrix formed at time index n as given in (6.19) and (6.20).

Alternatively, the CCM can be obtained as a Toeplitz matrix by stacking the individual cross-covariance matrices as in (6.26). These are Toeplitz matrices defined as

$$\mathbf{R}_{x,ml} = \begin{bmatrix} r_{x,ml}(0) & r_{x,ml}(1) & \dots & r_{x,ml}(P-1) \\ r_{x,ml}(-1) & r_{x,ml}(0) & \dots & \vdots \\ \vdots & & \ddots & \vdots \\ r_{x,ml}(-P+1) & \dots & & r_{x,ml}(0) \end{bmatrix} \tag{6.42}$$

where $r_{x,ml}(p)$ is the cross-correlation function between the signals at microphones m and l and is given by

$$r_{x,ml}(p) = \frac{1}{L}\sum_{n=0}^{L-1-|p|} x_m(n)x_l(n-p) \quad \text{for} \quad p = -P+1,\dots,P-1 \tag{6.43}$$

As in the single microphone case, the second approach seems more appropriate as it allows to preserve the Toeplitz structure of the covariance matrices which is found to yield better performance in speech enhancement applications [41]. This second approach is adopted in this thesis.

[8]The time index n is dropped in $\bar{\mathbf{R}}_x$ for simplicity.

6.6 The Eigen-Domain Averaging method

In this section we present a novel method for microphone array speech enhancement which is based on the MRQSS. This method exploits the structure of $\bar{\mathbf{R}}_s$ to modify the MRQSS using averaging in the eigendomain. This method is then referred to as the Multi-microphone Eigen-Domain Averaging method (MEDA).

6.6.1 Derivation

The MEDA is mainly based on the following important property which results from the special structure of the speech CCM $\bar{\mathbf{R}}_s$.

Property 6.6.1 *Suppose that* $\bar{\mathbf{v}}_i = [\mathbf{v}_{i1}^T, \ldots, \mathbf{v}_{iM}^T]^T$, *where* $\mathbf{v}_{im}$*'s* $(m = 1, \ldots, M)$ *are* P*-dimensional vectors, is the* i^{th} *unit norm eigenvector of* $\bar{\mathbf{R}}_s$ *with corresponding eigenvalue* $\bar{\lambda}_{s,i}$, *where* $\bar{\lambda}_{s,i} > 0$, *i.e. it is one of the first* Q *principal eigenvalues of* $\bar{\mathbf{R}}_s$. *Then we have*

$$\mathbf{v}_{i1} = \mathbf{v}_{i2} = \ldots = \mathbf{v}_{iM} \tag{6.44}$$

and $\mathbf{u}_i = \sqrt{M}\mathbf{v}_{i1}$, *is a unit norm eigenvector of* $\mathbf{R}_s$ *with corresponding eigenvalue* $\frac{\bar{\lambda}_{s,i}}{M}$.

Proof: Using (6.29), the product $\bar{\mathbf{R}}_s\bar{\mathbf{v}}_i$ can be written as

$$\bar{\mathbf{R}}_s\bar{\mathbf{v}}_i = \mathbf{C}\mathbf{R}_s\mathbf{C}^T\mathbf{v}_i = \mathbf{C}\left(\mathbf{R}_s\sum_{m=1}^{M}\mathbf{v}_{im}\right)$$

Now since $\bar{\mathbf{v}}_i$ is an eigenvector of $\bar{\mathbf{R}}_s$ then by definition we have $\bar{\mathbf{R}}_s\bar{\mathbf{v}}_i = \bar{\lambda}_{s,i}\bar{\mathbf{v}}_i$, so forcing a subvector-wise equality we get[9]

$$\mathbf{R}_s\sum_{m=1}^{M}\mathbf{v}_{im} = \bar{\lambda}_{s,i}\mathbf{v}_{il} \quad \text{for} \quad l = 1, \ldots, M. \tag{6.45}$$

The left-hand side of (6.45) is constant for all l, and since $\bar{\lambda}_{s,i} > 0$, then we have (6.44). Using this result (6.45) can be written as

$$\mathbf{R}_s\mathbf{v}_{i1} = \frac{\bar{\lambda}_{s,i}}{M}\mathbf{v}_{i1} \tag{6.46}$$

[9]Note that for a P-dimensional vector $\mathbf{x}$, $\mathbf{C}\mathbf{x} = [\mathbf{x}^T, \mathbf{x}^T, \ldots, \mathbf{x}^T]^T$.

Therefore $\mathbf{v}_{i1}$ and $\bar{\lambda}_{s,i}/M$ are an eigenvector and the corresponding eigenvalue of $\mathbf{R}_s$, respectively.

Now to find the unit norm eigenvector, we note that

$$\bar{\mathbf{v}}_i^T \bar{\mathbf{v}}_i = \sum_{m=1}^{M} \mathbf{v}_{im}^T \mathbf{v}_{im} = M \mathbf{v}_{i1}^T \mathbf{v}_{i1} \tag{6.47}$$

and since by definition $\bar{\mathbf{v}}_i^T \bar{\mathbf{v}}_i = 1$, then it can be seen that $\mathbf{u}_i = \sqrt{M}\mathbf{v}_{i1}$ is the i^{th} unit norm eigenvector of $\mathbf{R}_s$.

■

For the remaining $PM - Q$ eigenvectors, and since $\bar{\mathbf{R}}_s$ is non-negative definite, the corresponding eigenvalues will be constant and equal to zero. Therefore

$$\mathbf{R}_s \sum_{m=1}^{M} \mathbf{v}_{im} = 0 \quad \text{for} \quad i = Q+1, \ldots, PM. \tag{6.48}$$

Two cases can then arise

$$\begin{aligned} &\textstyle\sum_{m=1}^{M} \mathbf{v}_{im} \in \mathcal{N}(\mathbf{R}_s) \quad \text{(I)} \\ &\textstyle\sum_{m=1}^{M} \mathbf{v}_{im} = 0 \qquad\quad \text{(II)} \end{aligned} \tag{6.49}$$

where $\mathcal{N}(\mathbf{R}_s)$ is the null space of $\mathbf{R}_s$ (the noise subspace).

Property 6.6.1 tells us that the eigenvector matrix $\bar{\mathbf{V}}_1$ can be written in terms of the eigenvector matrix $\mathbf{U}_1$ of $\mathbf{R}_s$ as follows

$$\bar{\mathbf{V}}_1 = \frac{1}{\sqrt{M}} \mathbf{C}\mathbf{U}_1 \tag{6.50}$$

where multiplying every column of $\mathbf{U}_1$ by $\mathbf{C}$ from the left has the effect of stacking M copies of the same vector to form a PM-dimensional vector. The scaling is needed to obtain unit-norm eigenvectors.

Therefore, given a matrix $\bar{\mathbf{V}}_1$ satisfying property 6.6.1, it can be easily seen that an estimate of $\mathbf{U}_1$ can be obtained by taking the average of all sub-vectors in every column. This estimate can be written with the aid of the $\mathbf{C}$ matrix as follows

$$\hat{\mathbf{U}}_1 = \sqrt{M} \left[\frac{1}{M} \mathbf{C}^T \bar{\mathbf{V}}_1 \right] = \frac{1}{\sqrt{M}} \mathbf{C}^T \bar{\mathbf{V}}_1 \tag{6.51}$$

where this step constitutes the eigendomain averaging operation.

Using (6.50), we can replace $\bar{\mathbf{V}}_1$ in (6.35) so that the filter $\bar{\mathbf{H}} = \bar{\mathbf{V}}_1 \bar{\mathbf{G}} \bar{\mathbf{V}}_1^T$ be written as

$$\begin{aligned} \bar{\mathbf{H}} &= \left(\frac{1}{\sqrt{M}}\mathbf{C}\mathbf{U}_1\right) \bar{\mathbf{G}} \left(\frac{1}{\sqrt{M}}\mathbf{C}\mathbf{U}_1\right)^T \\ &= \frac{1}{M}\mathbf{C}\left(\mathbf{U}_1 \bar{\mathbf{G}} \mathbf{U}_1^T\right)\mathbf{C}^T \end{aligned}$$

Let $\mathbf{H} = \mathbf{U}_1 \bar{\mathbf{G}} \mathbf{U}_1^T$ be a $P \times P$ matrix, then we have

$$\bar{\mathbf{H}} = \frac{1}{M}\mathbf{C}\mathbf{H}\mathbf{C}^T. \tag{6.52}$$

That is, using the definition of the matrix $\mathbf{C}$, we have

$$\bar{\mathbf{H}} = \frac{1}{M}\begin{bmatrix} \mathbf{H} & \cdots & \mathbf{H} \\ \vdots & \ddots & \vdots \\ \mathbf{H} & \cdots & \mathbf{H} \end{bmatrix} \tag{6.53}$$

Using this result, the clean speech estimate in (6.37) can be written as

$$\begin{aligned} \hat{\mathbf{s}} &= \frac{1}{M}\mathbf{C}^T\left(\bar{\mathbf{H}}\bar{\mathbf{x}}\right) \\ &= \frac{1}{M}\mathbf{C}^T\left(\frac{1}{M}\mathbf{C}\mathbf{H}\mathbf{C}^T\bar{\mathbf{x}}\right) \end{aligned}$$

Recognizing $\mathbf{C}^T\mathbf{C}$ as $M\mathbf{I}_P$ and re-arranging terms, we get

$$\hat{\mathbf{s}} = \mathbf{H}\left(\frac{1}{M}\mathbf{C}^T\bar{\mathbf{x}}\right) \tag{6.54}$$

which can be viewed as a conventional beamformer with a signal subspace postfilter $\mathbf{H}$.

6.6.2 Handling spatio-temporal colored noise

Now we need to reformulate (6.38) in order to obtain the noise energies $\bar{\xi}_i$'s using the eigenvector matrix $\mathbf{U}_1$. Recalling the definition of $\bar{\xi}_i$ and using (6.50) we get

$$\begin{aligned} \bar{\xi}_i &= \bar{\mathbf{v}}_i^T \bar{\mathbf{R}}_w \bar{\mathbf{v}}_i \\ &= \left(\tfrac{1}{\sqrt{M}}\mathbf{C}\mathbf{u}_i\right)^T \bar{\mathbf{R}}_w \left(\tfrac{1}{\sqrt{M}}\mathbf{C}\mathbf{u}_i\right) \\ &= \mathbf{u}_i^T \underbrace{\left(\frac{1}{M}\mathbf{C}^T \bar{\mathbf{R}}_w \mathbf{C}\right)}_{\mathbf{A}} \mathbf{u}_i \end{aligned} \tag{6.55}$$

where the matrix $\mathbf{A}$ is actually a weighted sum of all the M^2 block matrices, $\mathbf{R}_{w,ml}$, of dimension $P \times P$ corresponding to the noise cross-covariance matrices between microphones m and l, i.e.

$$\mathbf{A} = \frac{1}{M}\mathbf{C}^T \bar{\mathbf{R}}_w \mathbf{C} = \frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{M} \mathbf{R}_{w,ml} \tag{6.56}$$

Therefore the matrix $\mathbf{A}$ is actually a weighted estimate of the noise covariance at the output of the fixed beamformer which can be shown to be equal to $\mathbf{A}/M$. Indeed, in the case of an incoherent noise field, i.e. $\mathbf{R}_{w,ml} = 0$ for $m \neq l$, $\mathbf{A}$ reduces to the noise covariance matrix $\mathbf{R}_w$, where it is assumed that the noise has the same covariance matrix over all microphones.

Now noting that $\bar{\lambda}_{s,i} = M\lambda_{s,i}$, the gain coefficients on every spectral direction defined in (6.39) can be re-written as

$$\bar{g}_i = \frac{\lambda_{s,i}}{\lambda_{s,i} + \mu\dfrac{\bar{\xi}_i}{M}} = \frac{\lambda_{s,i}}{\lambda_{s,i} + \mu\xi_i} = g_i \tag{6.57}$$

where $\xi_i = \bar{\xi}_i/M$, which suggests that the gain function actually accounts for the fact that the noise energy at the output of the beamformer would be reduced by a factor of M.

With the above described procedure the errors resulting from the approximation (6.40) are reduced especially under coherent noise fields where the MEDA method significantly outperforms the MRQSS method as will be shown in the experimental results. Indeed, while the Raleigh Quotient approach is found to be a reasonable approximation, though inaccurate mathematically, to handle temporal colored noise, it seems to be not really appropriate

to deal with coherent noise fields. This resulted in the relatively poor performance of the MRQSS particularly under directional noise. In MEDA, on the other hand, the averaging process carried out in (6.56) in addition to the eigendomain averaging carried out to estimate the eigenvectors, has the benefit of reducing the impact of noise hence improving the performance especially under diffuse fields. Under directional noise, this type of averaging helped to maintain a considerably better performance of MEDA over MRQSS. Actually an analogous approach is commonly used in frequency domain adaptive postfiltering methods to cancel the noise components as can be seen in (6.15).

6.6.3 Analysis

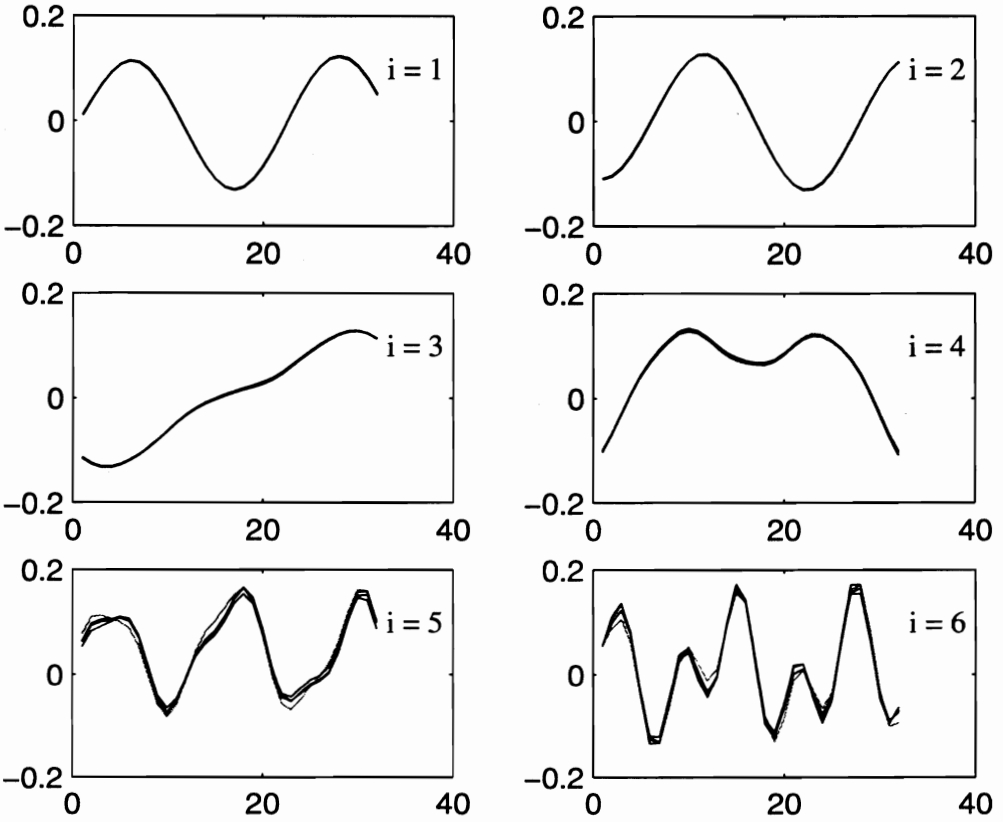


Fig. 6.3 The four subvectors of the first six eigenvectors of $\bar{\mathbf{R}}_s$ calculated during a frame containing the vowel /o/, for a signal corrupted by an incoherent white noise at 0 dB input SNR.

In this section we assess the validity of property 6.6.1 by examining the eigenvectors

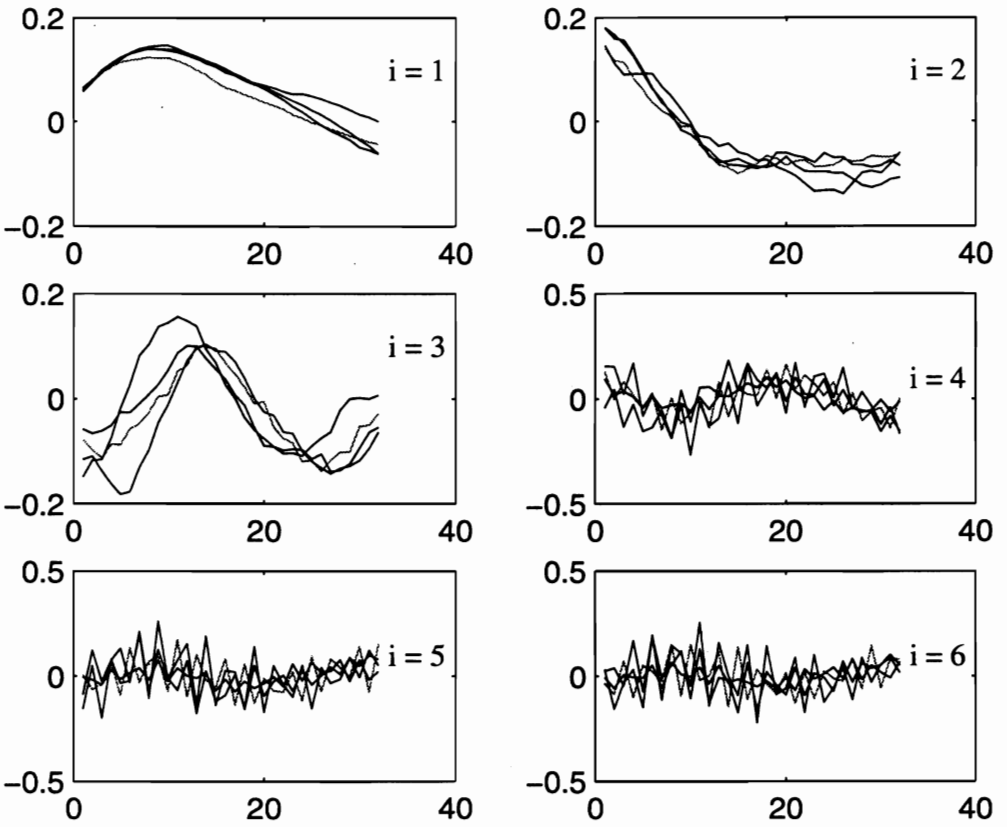


Fig. 6.4 The four subvectors of the first six eigenvectors of $\bar{\mathbf{R}}_s$ calculated during a frame containing the fricative /f/, for a signal corrupted by an incoherent white noise at 0 dB input SNR.

of the CCM $\bar{\mathbf{R}}_s$ and observing to what extent and under what conditions equation (6.44) holds.

To this end, we consider an array of $M = 4$ microphones in an incoherent noise field. Three different frames corresponding to a noise only frame, a fricative /f/ and a vowel /o/ are to be examined. These frames are picked from a sentence in which the desired signal is corrupted by computer generated white noise at 0 dB SNR. For every frame we plot the subvectors $\mathbf{v}_{im}$, for $m = 1, \ldots, M$ corresponding to the first six eigenvectors[10] $\bar{\mathbf{v}}_i$, for $i = 1, \ldots 6$, of $\bar{\mathbf{R}}_s$.

These waveforms are shown in Figures 6.3 , 6.4 and 6.5 for the vowel /o/, the fricative /f/ and the noise frame respectively. The norms of the eigenvector estimates $\hat{\mathbf{u}}_i$, obtained by taking the average of the subvectors as in (6.50), is shown in Figure 6.6 for the three

[10]These eigenvectors have the largest eigenvalues.

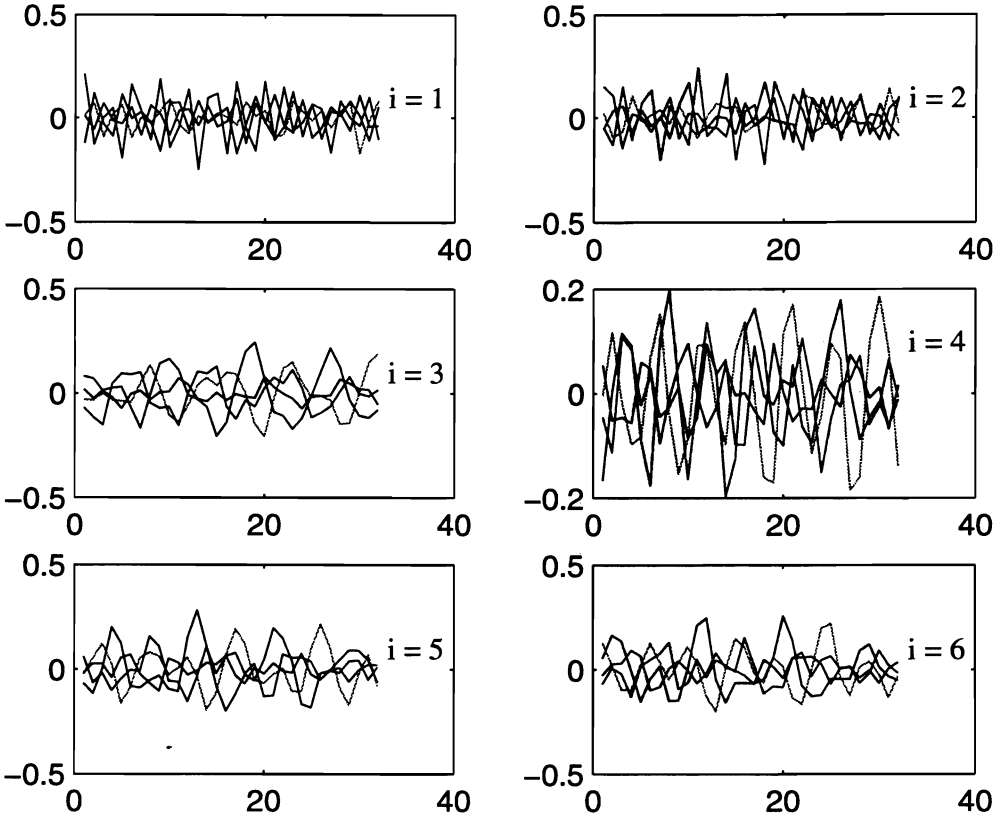
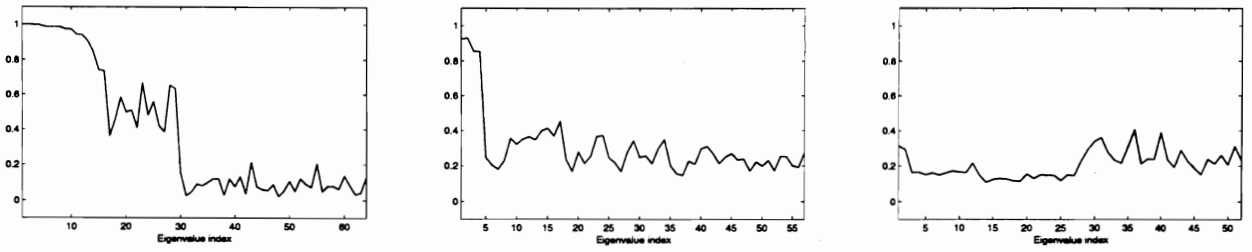


Fig. 6.5 The four subvectors of the first six eigenvectors of $\bar{\mathbf{R}}_s$ calculated during a noise only frame, for a signal corrupted by an incoherent white noise at 0 dB input SNR.

frames. Shown are the norms $\|\hat{\mathbf{u}}_i\|^2$ for the eigenvectors with strictly positive eigenvalues $\bar{\lambda}_{s,i}$, i.e. for $i = 1, \ldots, Q$.

For the vowel, it can be seen in Figure 6.3 that (6.44) is satisfied to a large extent for the shown eigenvectors, which usually carry the most important speech content. A slight deviation from the exact match can be seen in the 5^{th} and 6^{th} eigenvectors which will result in the norms $\|\hat{\mathbf{u}}_i\|^2$ being less than unity as can be seen in Figure 6.6 (a). This suggests that along these particular spectral directions the SNR is low enough for the noise to have an impact on the eigenvectors. This phenomenon can be more clearly observed in the fricative case shown in Figure 6.4. Indeed, this speech sound is usually weak with an SNR much lower than in a vowel frame. For this reason the discrepancies between the four subvectors can be seen even with the eigenvector with the largest eigenvalue (i.e. $i = 1$). Actually the averaging operation yielding the eigenvector estimate $\hat{\mathbf{u}}_i$ is expected to reduce the effect of



(a) Vowel /o/ (b) Fricative /f/ (c) Noise only

Fig. 6.6 The norm $\|\hat{\mathbf{u}}_i\|^2$ of the eigenvector estimates of $\mathbf{R}_s$ under incoherent white noise at 0 dB SNR, for three frames with different speech content.

noise hence improving the overall speech enhancement performance by obtaining a more reliable subspace estimate.

In the noise only frame results shown in Figure 6.5 it can be seen that the subvectors seem to have a largely random relationship with respect of each other making their average be relatively close to zero as can be seen in the norm $\|\hat{\mathbf{u}}_i\|^2$ for this particular frame shown in Figure 6.6 (c). It can be seen in that figure that the norm has an almost constant value equal to 0.2 for most of the eigenvalue indices. This complies with case (II) in (6.49) which states that the subvectors of $P(M-1)$ eigenvectors would actually add up to zero. The other norms which have a slight larger value may correspond to case (I) where the sum of the subvectors belongs to the null space of $\mathbf{R}_s$. This phenomenon would result in the cancellation of the (speech free) signal content in those directions, hence obtaining an improved noise reduction performance while introducing little or no distortion.

This result suggested the use of a thresholding mechanism in which the eigenvectors $\hat{\mathbf{u}}_i$'s with norms below a certain threshold are omitted from the matrix $\hat{\mathbf{U}}_1$. However, experimental simulations showed that the performance was highly dependent on the choice of the threshold and similar results could also be obtained by adjusting the control parameter μ in the gain function (6.57). Actually, for any eigenvector with a very low $\|\hat{\mathbf{u}}_i\|^2$, the corresponding noise energy ξ_i, as calculated in (6.55), will also be low and will approach zero. Therefore, and since the eigenvalue $\lambda_{s,i}$ is not affected by this phenomenon, the SNR in that direction will be very high leading to a gain coefficient close to one in (6.57). Thus, while the gain matrix will entail no noise reduction, the signal energy will be reduced in that spectral direction by the low norm of the eigenvector estimate $\hat{\mathbf{u}}_i$. Hence, instead

of relying on a pre-selected threshold, a better effect is actually automatically obtained in which the tradeoff between the gain coefficient and the eigenvector norm balances any signal subspace estimation errors.

Sensitivity to steering errors

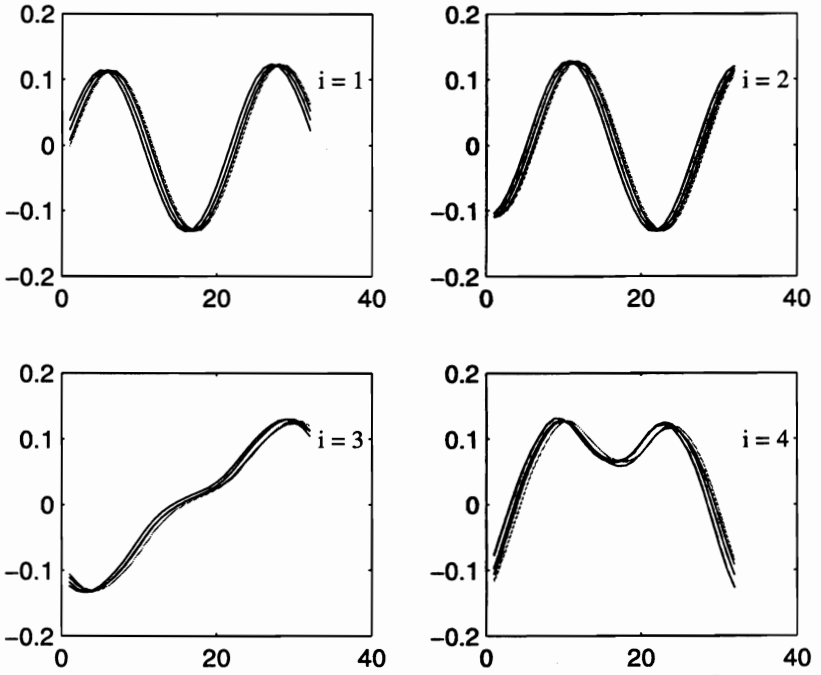


Fig. 6.7 The four subvectors of the first six eigenvectors of $\bar{\mathbf{R}}_s$ of the vowel /o/ extracted from a signal corrupted by an incoherent white noise at 0 dB input SNR. The desired speech DOA is 10 degrees.

In the following experiment we try to assess the effect of steering misadjustment that may result from errors in the time delay compensation module. In this experiment the same setup as described earlier is maintained except that now the desired speech signal impinges on the array at an angle of 10 degrees off the look direction (the array is steered to 0 degrees). Figures 6.8 and 6.7 show the waveforms of the subvectors $\mathbf{v}_{im}$, $m = 1, \ldots M$, for /o/ and /f/, respectively, for the first four eigenvectors[11], i.e. $i = 1, \ldots, 4$. Figure 6.9 shows the norms $\|\hat{\mathbf{u}}_i\|^2$ for $i = 1, \ldots, Q$.

[11]Note that while comparing Figures 6.3 and 6.7 can be instructive, comparing Figures 6.4 and 6.8 for the fricative /f/ is not possible since the order of the eigenvectors has changed due to the sorting operation and not all of them are illustrated.

It can be seen from these figures that the eigenvectors were indeed affected by this steering misadjustment. The effect can be better seen in Figure 6.7 where the subvectors, which had an almost exact match in Figure 6.3, seem now to have a constant phase shift with respect of each other. This phase shift resulted in a reduction of the norm of the eigenvector estimates $\|\hat{\mathbf{u}}_i\|^2$ as shown in Figure 6.9 (a). The effect on the fricative /f/ is higher since it is coupled with a higher vulnerability to noise, as discussed in the previous experiment.

This would indeed affect the performance by increasing the distortion due to the overall signal cancellation (including the noise). However, this drawback is not as serious as it appears to be because the overall effect of these steering errors is actually reduced by the trade off offered by the increased gain coefficients (because the estimated noise energies will be lower). This claim will be experimentally verified in Chapter 8.5.

Our experiments also revealed that the phase shift between the subvectors depended on the directional of arrival of the desired speech signal and on the inter-microphone distance. This suggests that if further research is conducted in this direction, this result may be used for time delay estimation leading to a self calibrating microphone array. The accuracy of such an approach and its ultimate performance has yet to be investigated.

6.6.4 The overall MEDA algorithm

In this section we present the overall detailed algorithm of the MEDA method.

To reduce the computational load, and as in the single microphone case, we use the frame-based SSA implementation developed in this thesis and presented in Section 5.2. Accordingly, the samples in every length L frame from the M channels are used to estimate the cross-correlation coefficients as in (6.43). Then, these coefficients are used to form the noisy CCM $\bar{\mathbf{R}}_x$ directly arranged into a Toeplitz structure as shown in (6.42) and (6.26). The noise CCM $\bar{\mathbf{R}}_w$ is obtained in a similar way during non-speech activity periods. The EVD of the difference $\bar{\mathbf{R}}_x - \bar{\mathbf{R}}_w$ is used to obtain an estimate of the clean speech CCM eigenvector and eigenvalue matrices $\bar{\mathbf{V}}$ and $\bar{\mathbf{\Lambda}}_s$ respectively.

The rank Q of $\bar{\mathbf{R}}_s$ is obtained as the number of strictly positive eigenvalues $\bar{\lambda}_{s,i}$. The corresponding eigenvectors are used to form the columns of the transformation matrix

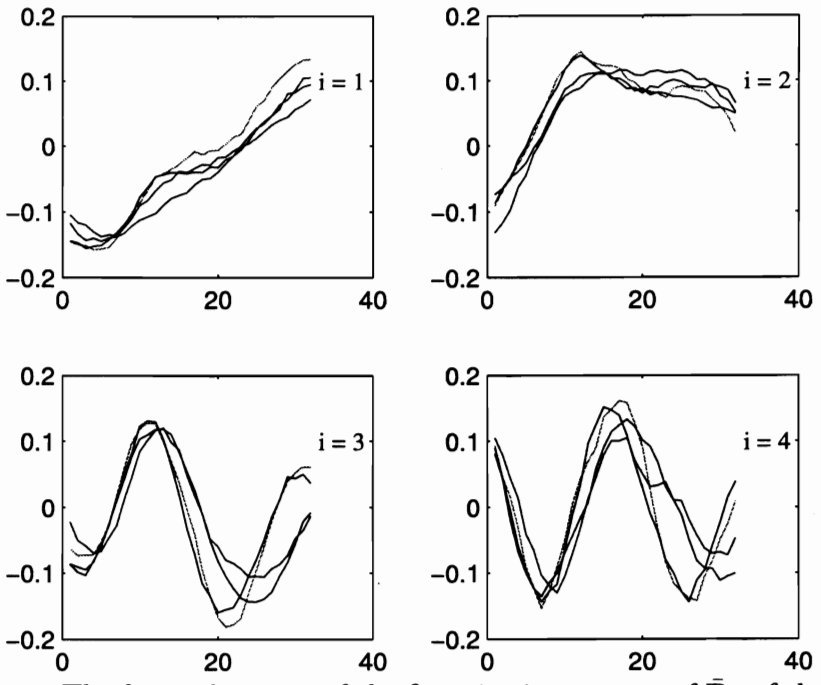


Fig. 6.8 The four subvectors of the first six eigenvectors of $\bar{\mathbf{R}}_s$ of the fricative /f/ extracted from a signal corrupted by an incoherent white noise at 0 dB input SNR. the desired speech DOA is 10 degrees.

$\bar{\mathbf{V}}_1 = [\bar{\mathbf{v}}_1, \bar{\mathbf{v}}_2 \ldots, \bar{\mathbf{v}}_Q]$. The matrix $\bar{\mathbf{V}}_1$ is used to estimate the matrix $\mathbf{U}_1$ as follows

$$\hat{\mathbf{U}}_1 = \frac{1}{\sqrt{M}} \mathbf{C}^T \bar{\mathbf{V}}_1 = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_Q] \tag{6.58}$$

the noise energies are then calculated as

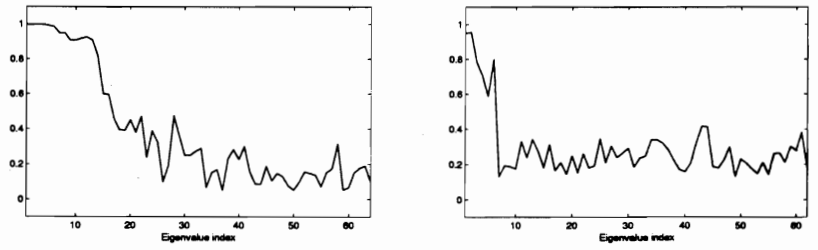
$$\xi_i = \frac{\bar{\xi}_i}{M} = \frac{1}{M^2} \hat{\mathbf{u}}_1^T \left(\mathbf{C}^T \bar{\mathbf{R}}_w \mathbf{C} \right) \hat{\mathbf{u}}_i, \quad \text{for } i = 1, \ldots, Q \tag{6.59}$$

and the speech eigenvalues are obtained as

$$\lambda_{s,i} = \frac{\bar{\lambda}_{s,i}}{M}, \quad \text{for } i = 1, \ldots, Q \tag{6.60}$$

Using these quantities, the entries of the gain matrix $\mathbf{G} = [g_1, \ldots, g_Q]$ are calculated as follows

$$g_i = \frac{\lambda_{s,i}}{\lambda_{s,i} + \mu \xi_i} \tag{6.61}$$



(a) Vowel /o/ (b) Fricative /f/

Fig. 6.9 The norm $\|\hat{\mathbf{u}}_i\|^2$ of the eigenvector estimates of $\mathbf{R}_s$ under incoherent white noise at 0 dB SNR, for two frames with different speech content. In this case the desired speech signal DOA is 10 degrees.

The clean speech estimate is finally obtained as

$$\hat{\mathbf{s}} = \hat{\mathbf{U}}_1 \mathbf{G} \hat{\mathbf{U}}_1^T \left(\frac{1}{M} \mathbf{C}^T \mathbf{x} \right) \tag{6.62}$$

The individual overlapping P-dimensional vectors estimated as in (6.62) are combined using the overlap-add technique and the total speech signal is obtained by combining all frames as explained in Section 5.2. Note that the matrix $\mathbf{C}$ is used here just to simplify the notation and that the resulting multiplications are actually avoided by performing the corresponding equivalent additions as explained in Appendix A.

6.7 Including the perceptual criteria

Since the MEDA subspace filter is applied as a postfilter, it is straight forward to couple this method with the Perceptual Signal Subspace (PSS) method presented in Chapter 5 in order to improve the overall performance by including the human hearing properties in the design. This combination can be done by using MEDA to estimate the eigenvalues and the eigenvector matrix of the clean speech covariance matrix $\mathbf{R}_s$. Then, the gain coefficients in (6.57) are modified to include the perceptual criteria as described in Section 5.6.

6.8 Summary

In this chapter we have presented a novel multi-microphone method based on the signal subspace approach. The new method exploits a property of the EVD of the signal composite covariance matrix in order to perform averaging in the eigendomain, hence the name MEDA. This property states that an eigenvector of the CCM has equal subvectors. The weighted sum of those subvectors can then be used to estimate the eigenvectors of the speech covariance matrix which span its signal subspace.

Chapter 7

The subband room response simulator

To be able to design and evaluate the microphone array methods presented in the thesis, it is useful to have a fast and flexible way to digitally simulate the effect of a reverberant room on the acoustic signals at the input of the microphone array. To this end, the image method [2], was proposed to simulate the discrete-time room impulse response. The contribution of every image to the total effect, is represented by a weighted impulse shifted to the discrete time instant closest to the actual arrival time.

The image method, however, does not precisely estimate the echo arrival times because the latter are usually not multiples of the sampling period [126]. Hence, in multi-microphone systems for example, this method fails to give a good estimate to the very important inter-microphone phase resulting in serious simulation errors. To solve this problem, Peterson suggests to distribute the arriving echo over several samples according to a low pass response function centered at the actual echo arrival time [126]. A generalization of this method to a moving point source in a reverberant room is presented in [17].

Peterson's modified image method, although being widely used, is not capable of providing a realistic estimate of the room impulse response. The reason is that it does not account for the dependency of the wall reflection coefficients on frequency. Instead, the low-pass image method considers that the reflection coefficients are constant over all the frequency range of interest. Table 7.1 (from [132]) shows the frequency dependent absorption coefficients (equal to one minus the reflection coefficients squared) for some common

material used in rooms.

Table 7.1 A brief list of absorption coefficients of some materials, as a function of frequency in Hz [132].

	Frequency (Hz)					
Material	**125**	**250**	**500**	**1000**	**2000**	**4000**
Concrete block, unpainted	0.36	0.44	0.31	0.29	0.39	0.25
Concrete block, painted	0.10	0.05	0.06	0.07	0.09	0.08
Glass, window	0.35	0.25	0.18	0.12	0.07	0.04
Plaster on lath	0.14	0.10	0.06	0.05	0.04	0.03
Plywood paneling	0.28	0.22	0.17	0.09	0.10	0.11
Carpet on pad	0.08	0.24	0.57	0.69	0.71	0.73
Gypsum board, one-half inch	0.29	0.10	0.05	0.04	0.07	0.09
Drapery, lightweight	0.03	0.04	0.11	0.17	0.24	0.35

Another difficulty with the image method is the computation cost of the convolution operation needed to calculate the acoustic response between two points in a reverberant room. Precisely, L multiplications per input sample, where L is the length of the impulse response, are required. Unfortunately, because of the room acoustic properties, L is generally a large number. This computational burden increases in the multi-microphone case because the costly convolution is repeated with all microphones.

In this chapter we present a subband room simulator (SRS) with added design flexibility for modeling the frequency dependent reflection coefficients in a computationally efficient manner. This is achieved using a subband scheme where the input signal is divided into K channels and every subband signal is convolved with a subband impulse response (SIR) at a sampling rate reduced by a factor $M \leq K$. We first show how to calculate the SIR's so that if the reflection coefficients are kept constant with respect to frequency, then the resultant overall impulse response is equal to that designed with the low-pass image method. We then show how the SIR's can be accordingly modified so that an overall impulse response with frequency dependent reflection coefficients is achieved. In addition to this added design flexibility, the implementation of the SRS significantly reduces the number of required multiplications compared to the image method thereby providing an efficient solution to the room impulse response simulation problem.

7.1 Uniform DFT filter banks

Subband filtering has been used efficiently in various applications in the field of signal processing especially in speech coding and adaptive filtering [152, 24]. It owes its popularity to the frequency flexibility and the computational savings that it offers.

In our approach, we exploit these two properties to provide a fast implementation scheme to the image method and to provide more degrees of freedom to the currently used image method.

In the proposed SRS, a uniform DFT filter bank is used to realize the subband analysis and synthesis. In this approach, the input signal is divided into K adjacent subbands by a bank of complex demodulators whose outputs are lowpass filtered by the antialiasing filter $f(n)$ and then downsampled by M to a lower sampling rate. After performing any desired modification on the subband signals, they are upsampled to the initial rate, lowpass filtered by the anti-imaging filter $g(n)$ and modulated back to their original spectral position. The output is finally obtained by summing the sub-signals. A critical design issue here is the choice of the downsampling factor M. In many practical implementations oversampling, that is $M < K$, is used. It is indeed one of the simplest and most efficient ways to reduce the subband aliasing when modifications are made to the subband signals [49]. Several methods are now available for the design and realization of oversampled DFT filter banks [29, 63]. In this chapter we use the approach of [110] for the filterbank design which also proved useful in the context of acoustic echo cancellation. A detailed analysis of this filter bank design method can be found in [110]; however for convenience of our upcoming derivations, we present a brief summary of it here. The block diagram of the specific DFT filterbank under consideration is illustrated in Figure 7.1, where $x(n)$ denotes the input signal sampled at the rate F_s.

The z-transform $X_k(z)$ of the subband signals $x_k(m)$, where $k = 0 \ldots K-1$ denotes the subband index and m is the discrete time index at low sampling rate (i.e. F_s/M), is written as

$$X_k(z) = \frac{1}{M} \sum_{m=0}^{M-1} F(z^{1/M} W_M^{-m}) X(z^{1/M} W_M^{-m} W_K^k) \tag{7.1}$$

where $W_K = \exp(j2\pi/K)$ and $W_M = \exp(j2\pi/M)$. $F(z)$ and $G(z)$ are the z-transforms of the filters $f(n)$ and $g(n)$ respectively. In the z-domain, the synthesizer output $y(n)$ may be

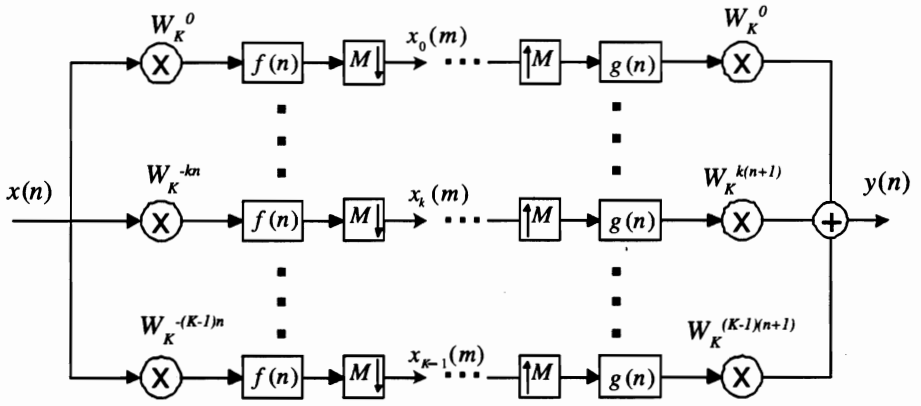


Fig. 7.1 Block diagram of the DFT filterbank.

expressed as

$$Y(z) = \frac{1}{M} \sum_{m=0}^{M-1} T_m(z) X(z W_M^{-m}) \tag{7.2}$$

The approach of [110] uses a special set of modulators, namely: W_K^{-kn} in the analysis and $W_K^{k(n+1)}$ in the synthesis. Assuming for now that the subband signals are not modified, we have

$$T_m(z) = \sum_{k=0}^{K-1} W_K^k F(z W_K^{-k} W_M^{-m}) G(z W_K^{-k}) \tag{7.3}$$

Because $F(z)$ and $G(z)$ have the same cutoff frequency ($\omega_c = \pi/K$), and since $M < K$, $T_m(z) \approx 0$ for $m \neq 0$ [110]. So the aliasing components can be neglected in (7.2) which reduces to

$$Y(z) = \frac{1}{M} T_0(z) X(z) \tag{7.4}$$

In order to have a constant group delay and hence reduce the phase distortion, $f(n)$ and $g(n)$ are chosen to be FIR filters of length N such that $g(n) = f(N - n - 1)$. Now if $N = n_f K$, where n_f is an integer, then $T_m(z)$, for $m = 0$, can be written in the frequency domain as

$$T_0(e^{j\omega}) = e^{-j(N-1)\omega} \sum_{k=0}^{K-1} \mid F(e^{j\omega} W_K^{-k}) \mid^2 \tag{7.5}$$

Obviously the response of the filter bank has a linear phase. Thus, by designing a prototype

filter $f(n)$ such that the magnitude of $T_0(e^{j\omega})$ is equal to one, i.e.

$$\sum_{k=0}^{K-1} \mid F(e^{j\omega} W_K^{-k}) \mid^2 \approx 1 \tag{7.6}$$

the total distortion will be a pure delay. In [110], the design of $f(n)$ to approximate these requirements is achieved by interpolation of 2-channel QMF filters (found in lookup tables [24]) by a factor of $K/2$.

For the implementation of the oversampled DFT filter bank, we use the *weighted-overlap-add* (WOA) structure because it allows flexibility in choosing M as any arbitrary integer [24]. This approach needs in the order of $2(\log_2 K + n_f)K/M$ multiplications per input sample (MPIS), which is generally a very low cost for most practical values of K, M and n_f.

7.2 The Subband Room Simulator (SRS)

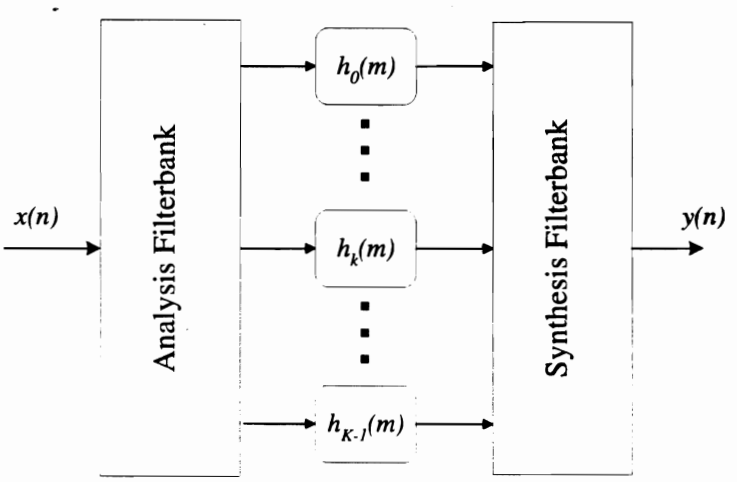


Fig. 7.2 Block diagram of the SRS algorithm.

The room simulator that we propose in this chapter is based on a subband implementation of the convolution with the room impulse response in the subbands at the low sampling rate. As seen in Figure 7.2, in which a block diagram of the SRS is shown, every subband signal is convoloved separately with a different subband impulse response (SIR). Calculating these SIR's is a critical design problem in our method. To this end, we first show how to

calculate the SIR's from a desired full band reference impulse response $h(n)$. The latter can be obtained from real measurements or can be a synthetic one calculated using the lowpass image method of [126], for instance. This method, just exploits the computational saving advantage of the SRS quantified in Section 7.2.3. However, it is an essential step in our developments because it explains, from a signal processing perspective, the reason behind the solution we are proposing for the subband impulse responses. After that, equipped with this solution, we present the subband image method which uses the image method to calculate $h_k(m)$ in a subband scheme.

7.2.1 Calculating the subband impulse responses

In this section, we show how to calculate the SIR $h_k(m)$, so that the system in Figure 7.2, has an impulse response equal to a desired room response $h(n)$. Including the convolution with $h_k(m)$, $T_m(z)$ in (7.3) becomes

$$T_m(z) = \sum_{k=0}^{K-1} W_K^k F(zW_K^{-k}W_M^{-m}) G(zW_K^{-k}) H_k(z^M W_K^{-kM}) \tag{7.7}$$

where $H_k(z)$ is the z-transform of $h_k(m)$. Our aim is to find an $H_k(z)$ which if upsampled and modulated (due to applying the synthesis bank to it) makes the transfer function of the whole system equal to the desired transfer function $H(z)$. As we explain below, this can be done by feeding $h(n)$ to an analysis bank so that

$$H_k(z) = \frac{1}{M} \sum_{m=0}^{M-1} \tilde{F}(z^{1/M} W_M^{-m}) H(z^{1/M} W_M^{-m} W_K^k) \tag{7.8}$$

where $\tilde{F}(z)$ is the z-transform of $\tilde{f}(n)$ which is an FIR lowpass filter with cutoff frequency $\gamma\pi/K$. Since $M < K$ then we can still argue that $T_m(z) \approx 0$ for $m \neq 0$, that is, and after substituting (7.8) in (7.7), the only term that remains is

$$T_0(z) = \sum_{k=0}^{K-1} W_K^k F(zW_K^{-k}) G(zW_K^{-k}) \frac{1}{M} \sum_{l=0}^{M-1} \tilde{F}(zW_K^{-k}W_M^{-l}) H(zW_M^{-l}) \tag{7.9}$$

If in addition we have $\gamma < K/M$ then we can and again argue that the pass-band of $F(zW_K^{-k})$ will fall well into the stop band of $\tilde{F}(zW_K^{-k}W_M^{-l})$, hence only the terms for $l = 0$

will remain in the second summation of the above equation. Thus $T_0(z)$ can be written as

$$T_0(z) = \frac{1}{M}\sum_{k=0}^{K-1} W_K^k F(zW_K^{-k})G(zW_K^{-k})\tilde{F}(zW_K^{-k})H(z) \tag{7.10}$$

Substituting (7.10) in (7.4) we get

$$Y(z) = \frac{1}{M^2}A(z)H(z)X(z) \tag{7.11}$$

where in the frequency domain $A(e^{j\omega})$ is

$$A(e^{j\omega}) = e^{-j(N-1)\omega}\sum_{k=0}^{K-1} |F(e^{j\omega}W_K^{-k})|^2\tilde{F}(e^{j\omega}W_K^{-k}) \tag{7.12}$$

Our experiments show that if $\tilde{f}(n) = f(n)$, then the magnitude of $A(e^{j\omega})$ will have valleys at the boundaries of the frequency subbands which significantly distort the final result. Therefore, $\tilde{f}(n)$ is chosen to have a slightly bigger bandwidth than $f(n)$ (that is $\gamma > 1$).

Property 7.2.1 *If $\tilde{f}(n)$ is chosen to be a linear phase FIR filter (designed using a window of length $N = n_f K$), and if n_f is an even number, then $A(e^{j\omega})$ will have a linear phase with group delay $d = \frac{3N}{2} - 1$ and its magnitude be given by*

$$|A(e^{j\omega})| = \sum_{k=0}^{K-1} |F(e^{j\omega}W_K^{-k})|^2|\tilde{F}(e^{j\omega}W_K^{-k})|. \tag{7.13}$$

Proof:

To begin the proof, we note that $\tilde{f}(n)$ has a group delay $p = N/2 = n_f K/2$, so

$$\tilde{F}(e^{j\omega}) = |\tilde{F}(e^{j\omega})|e^{-jp\omega} \tag{7.14}$$

Therefore

$$A(e^{j\omega}) = e^{-j(N-1)\omega} \sum_{k=0}^{K-1} |F(e^{j\omega}W_K^{-k})|^2 |\tilde{F}(e^{j\omega}W_K^{-k})| e^{-j(\omega-\frac{2\pi}{K}k)p} \quad (7.15)$$

$$= e^{-j(N-1+p)\omega} \sum_{k=0}^{K-1} |F(e^{j\omega}W_K^{-k})|^2 |\tilde{F}(e^{j\omega}W_K^{-k})| e^{j\frac{2\pi}{K}kp} \quad (7.16)$$

Using the fact that n_f is a multiple of 2, the exponential term in the summation can be written as

$$\exp\left(j\frac{2\pi}{K}kp\right) = \exp\left(j\frac{2\pi}{K}.\frac{n_f K}{2}.k\right) = \exp\left(jn_f k\pi\right) = 1 \quad (7.17)$$

Which completes the proof.

■

Since $A(e^{j\omega})$ has a linear phase then we are sure that there will be no phase distortion in the output signal $y(n)$ in (7.11). In what follows we show that with this choice of filters the magnitude distortion is at an acceptable level.

Since the bandwidth of $f(n)$ is $2\pi/K$, then just two consecutive terms of the summation in (7.13) need to be considered in our analysis, namely

$$|A(e^{j\omega})| = |F(e^{j\omega}W_K^{-k})|^2 |\tilde{F}(e^{j\omega}W_K^{-k})| + |F(e^{j\omega}W_K^{-(k+1)})|^2 |\tilde{F}(e^{j\omega}W_K^{-(k+1)})| \quad (7.18)$$

for $\omega \in [2\pi k/K, 2\pi(k+1)/K]$. If $\tilde{f}(n) = f(n)$ then at frequencies far enough from the subband edges, the magnitude of $A(e^{j\omega})$ will be close to unity as a consequence of (7.6). That is $|A(e^{j\omega})|$ will be close to the response of the filter-bank without the convolution operation. However, at $\omega = \frac{\pi k}{K}$, $f(n)$ is designed such that $|F(\pi/K)| = -3$ dB (at the cutoff frequency) [24], therefore, if $\tilde{f}(n) = f(n)$, then

$$|A(e^{j\pi k/K})| \approx 2(0.707)^3 \approx -3 \text{ dB} \quad (7.19)$$

This would result in large magnitude distortions at the subband boundaries. Hence in our design we choose to use a different filter $\tilde{f}(n)$ with a larger pass band. Namely, we use another lowpass filter $\tilde{f}(n)$, with the same length as $f(n)$, designed, for example, using a Hamming window. Far enough from the subband boundaries, $|A(e^{j\omega})|$ will still have an acceptable "flatness". At $\omega = k\pi/K$, on the other hand, $|A(e^{j\pi k/K})| \approx$ -6 dB when $\gamma = 1$.

Therefore to improve the performance at these frequencies, a filter with a larger cutoff frequency is used. It is difficult to quantify the improvement we obtain with this method but from Figure 7.3 it can be seen that increasing γ (without exceeding K/M) decreases the magnitude distortion from -6 dB at $\gamma = 1$ to -1.3 at $\gamma = 1.2$ and to as little as -0.4 dB at $\gamma = 1.3$.

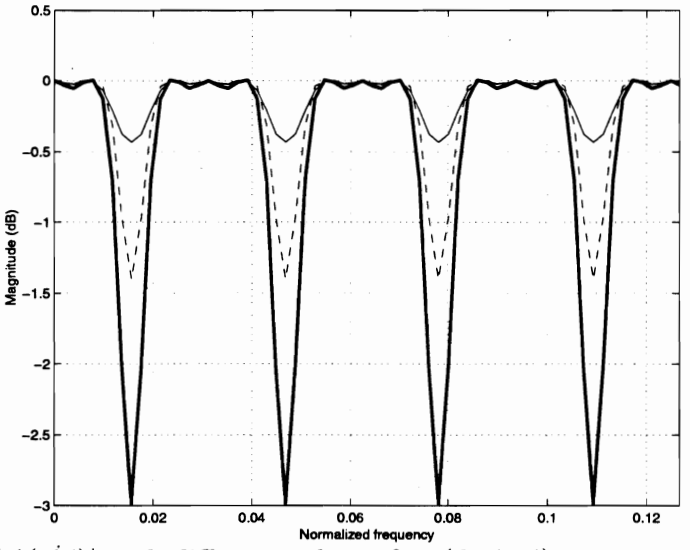


Fig. 7.3 $|A(e^{j\omega})|$ with different values of γ, (dashed) $\gamma = 1.2$ and (continuous) $\gamma = 1.3$, (thick) $\tilde{f}(n) = f(n)$ (because $|A(e^{j\omega})|$ is periodic, just a portion of the spectrum is shown for clarity).

7.2.2 The subband image method

Based on the results of the previous section, we can now derive the subband image method. In the lowpass impulse method [2, 126], the response $h(n)$ of a microphone at location x to an impulse excitation at location x_0 can be expressed as follows

$$h(n) = \frac{1}{4\pi} \sum_r \frac{\beta_r}{\tau_r} \psi_r(n) \tag{7.20}$$

where c is the speed of sound, r is the index of the image at position x_r, $\tau_r = \| x_r - x \| / c$ is the echo arrival of the image r, β_r is the corresponding composite reflection coefficient and $\psi_r(n)$ is a sampled version of a continuous, Hanning-windowed, lowpass filter $\psi_r(t)$

centered at the echo arrival time τ_r and is given by [126]

$$\psi_r(t) = \psi(t - \tau_r) \tag{7.21}$$

and

$$\psi(t) = \frac{1}{2}[1 + \cos(2\pi t/T_w)]\text{sinc}(2\pi f_c t) \qquad \text{for } -T_w/2 \leq t \leq T_w/2 \tag{7.22}$$

where T_w is the length of the lowpass filter and f_c is its cutoff frequency.

Now (7.8) written in the time domain becomes

$$h_k(m) = \sum_n \tilde{f}(mM - n)h(n)W_K^{-kn} \tag{7.23}$$

Then, substituting (7.20) in (7.23) and after changing the summation order we get

$$h_k(m) = \frac{1}{4\pi}\sum_r \frac{\beta_{r,k}}{\tau_r}\,\xi_{r,k}(m) \quad k = 0 \ldots K-1 \tag{7.24}$$

where

$$\xi_{r,k}(m) = \sum_n \tilde{f}(mM - n)\psi_r(n)W_K^{-kn} \tag{7.25}$$

Note that a subscript k is added to the composite reflection coefficient β_r to indicate that with this formulation it is possible to assign different reflection properties to each subband.

One possible way to implement (7.24) and (7.25) is to perform subband analysis to the lowpass function $\psi_r(n)$ corresponding to every new image and update the subband responses $h_k(m)$ accordingly. Once the SIR's $h_k(m)$ are computed they are used as in Figure 7.2 to simulate a synthetic room impulse response.

7.2.3 Computational load

At this reduced sampling rate, the SIR's $h_k(m)$ have a length L/M where L is the length of the overall impulse response $h(n)$. Consequently, L/M multiplications are needed every M samples for all the K channels. So, KL/M^2 MPIS are required to accomplish the convolution operations. On the other hand, if the putput is obtained by direct convolution with $h(n)$, L MPIS will be needed. Hence the total computational savings will be in the

order of

$$\frac{M^2}{K}\left[\frac{1}{1+\alpha(L)}\right] \tag{7.26}$$

where $\alpha(L) = 2M(\log_2 K + n_f)/L$. For most practical cases, L is very large so $\alpha(L)$ is usually a fraction between 0 and 1. Asymptotically, $\alpha(L) \to 0$ so the savings are in the order of M^2/K.

Actually since the subband signals are complex, the convolution requires twice as much multiplications as stated above. However, since both the input signal $x(n)$ and the room impulse response $h(n)$ are real in the present application, this can be overcome using the symmetry properties of the subband signals and avoid unnecessary repeated calculations. That is, $Y_k(m) = Y^*_{K-k}(m)$ for $k = 0 \ldots K/2 - 1$.

7.3 Experimental Results

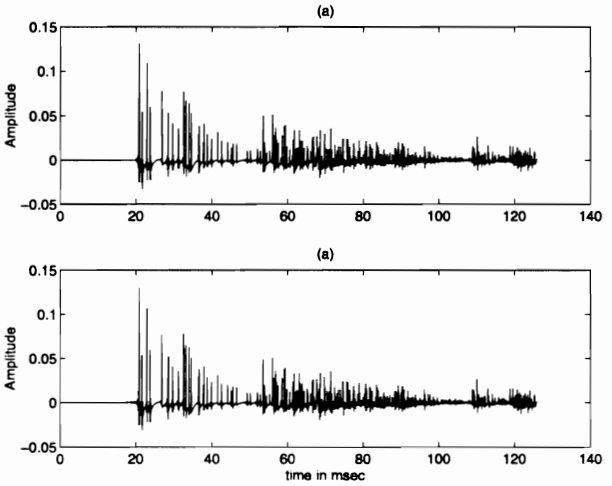


Fig. 7.4 Image method impulse response $h(n)$ (up) SRS impulse response $\hat{h}(n+d)$ (down).

In this section we describe the experiments made to test the performance of the new subband room simulator. We start by comparing the simulated room impulse response of the new method with that of the original image method, so the reflection coefficients in this case are constant for all frequencies. The room to be simulated has dimensions (15,10,4) with a microphone at positions (9,3.75,0.7) and a loudspeaker at (2,1.5,1.5). The dimensions are in meters and the origin of the coordinate system is at one of the lower

corners of the room. The reflection coefficients are set to 0.9 for the walls and 0.7 for the floor and ceiling. The length of the impulse response is 256 msec that is $L = 2048$ samples at 16 KHz sampling rate.

For the filter bank we use the following parameters, $K = 32$, $M = 24$, $N = 256$. The prototype filter $f(n)$ is obtained via interpolation (see [110]) from the 2-channel QMF filter 16A in [24] ($n_f = 8$). The second filter $\tilde{f}(n)$ is designed by windowing (using a Hamming window) with $\gamma = 1.3$. With these parameters $\alpha(L)$ in (7.26) is 0.3 leading to savings in the number of multiplications by a factor of about 13 times. The impulse response $h(n)$

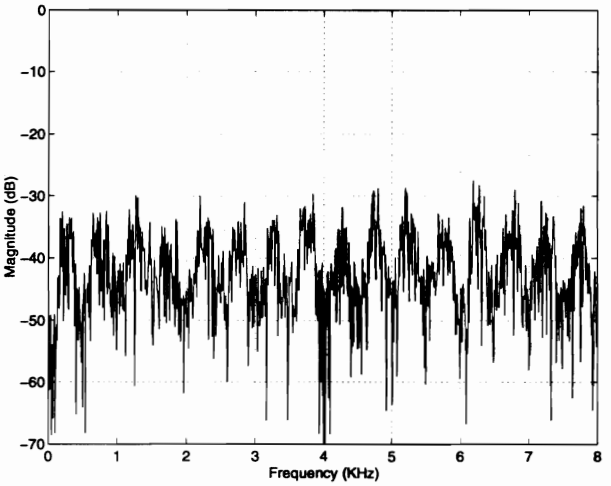


Fig. 7.5 Magnitude Squared of the error $|E(e^{j\omega})|^2$

computed with the image method [2, 126], and the SRS impulse response $\hat{h}(n+d)$ (where d is the group delay of $A(e^{j\omega})$) computed with the direct method of Section 3.1, are shown in Figure 7.4. Figure 7.5 illustrates the magnitude squared of the Fourier transform $E(e^{j\omega})$ of the error function $e(n) = h(n) - \hat{h}(n+d)$. The arithmetic mean of the error squared is -40 dB. Figure 7.6 shows how this arithmetic mean varies with the choice of the downsampling factor M when $K = 32$. This figure justifies our choice for $M = 24$.

Additional experiments were made with audio data of speech, music and percussion recordings. A few persons were asked to listen to the recordings after adding reverberation to them using full band convolution with $h(n)$ and using the SRS. To have some psychological effect, these people were indirectly given the impression that their ability to detect some kind of difference between two audio recordings is being tested. However, they disappointedly reported their inability to perceive any difference.

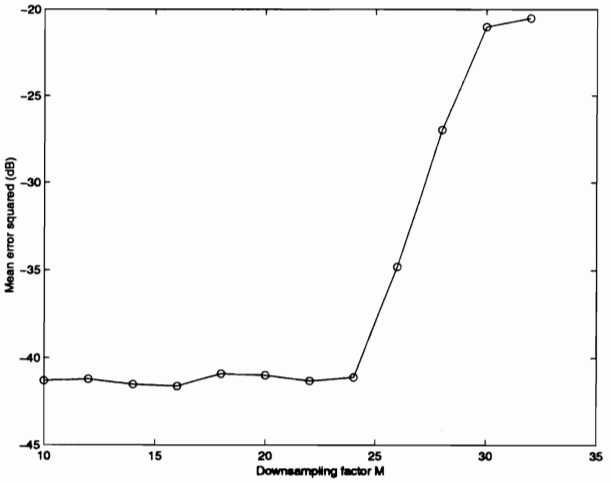


Fig. 7.6 Mean error squared in dB versus the downsampling factor M

Further simulations were made using the subband image method of Section 7.2.2 to test the effect of varying the reflection coefficients with frequency. Demo files with description of the room environment of every experiment can be found on the web site [140]. These demo files demonstrate the flexibility so achieved to simulate different environments, which is not possible with the classical image method. With this method acoustic properties close to those of very reverberant enclosures like churches and Turkish baths were simulated.

7.4 Conclusion

In this chapter, a fast Subband Room Simulator (SRS) is presented. Basically the method uses a subband filter bank to perform convolution operations at a reduced sampling rate, hence reducing the computational complexity. This reduction facilitates the evaluation of algorithms designed for the increasingly growing research area of multi-microphone systems. The interesting property of the SRS is that it can be implemented in a way that offers more flexibility in choosing the room acoustic parameters. Namely, different reflection coefficients can be assigned to different frequency bands. Our experiments show that no difference can be perceived between reverberation added using the traditional image method or using the SRS with constant reflection coefficients. Moreover experimental results show that changing the reflection coefficients over the frequency range of interests allows to simulate more realistic acoustic environments.

Chapter 8

Experimental Results

In this chapter we present experimental results evaluating the different novel signal subspace techniques developed in this thesis. We first begin the chapter by describing the performance evaluation tools utilized and then we introduce the setup employed in the various experiments.

In Section 8.3, we evaluate the novel fast Frame Based EVD (FBEVD) implementation scheme we propose for the signal subspace approach. In this technique, presented in Section 5.2, the stationarity of the speech signal is exploited in order to reduce the rate at which the signal subspace filter is updated. Consequently, the costly EVD calculation becomes less frequent resulting in considerable computational savings. The experiments carried out first measure any incurred signal degradation then the computational savings are quantified.

In Section 8.4, we assess the performance of the new Perceptual Signal Subspace method (PSS) introduced in Chapter 5. In this method, human masking properties represented by a threshold calculated by means of a frequency domain masking model, are mapped to the eigendomain via a Frequency to Eigendomain Transformation (FET). Doing so, a set of masking energies are acquired and used to modify the signal subspace gain function and have it bear perceptual criteria. In the experiments, we present some informal listening tests observations supported by spectrogram illustrations. Then, the results of several subjective tests which measure the performance of PSS against competing signal subspace methods are given.

Finally, in Section 8.5, the novel Multi-microphone signal subspace method with Eigendomain Averaging (MEDA) is evaluated. This approach, described in Chapter 6, takes

advantage of the structure of the spatio-temporal covariance matrix, referred to as the composite covariance matrix (CCM), in order to perform averaging in the eigendomain. This operation contributes to cope with the corrupting noise interference hence results in robust signal subspace filters. MEDA is evaluated against competing methods under different SNR levels and different reverberation conditions for various noise types. The sensitivity of MEDA to steering errors is also assessed.

8.1 Performance measures

It is important for all speech enhancement methods to have available tools which measure their performance and provide a means to compare them against other competing methods.

One such tool is to assess the performance visually via graphical illustrations. Waveform illustrations of the speech signals, though can be sometimes useful, are not usually very informative. Spectrograms, on the other hand, can offer a more valuable performance evaluation since more accurate conclusions about the residual noise level (and shape) and the signal distortion can be drawn from them. In this Chapter, spectrograms are used for evaluation and they will be often complemented by supporting observations from informal listening tests.

Actually, the most reliable measure in the context of speech, is based on subjective evaluation. In subjective tests, human listeners (the actual and eventual users of the tested methods) are asked to give their opinion on the enhanced signals based on some specified criteria. Subjective measures will be used in this Chapter to evaluate the performance of the proposed PSS method.

Subjective tests are usually very time consuming and difficult to make because the human subjects who take the tests are not usually available as desired. Therefore, alternatively, objective measures are often used. A thorough survey of different objective measures and their correlation with subjective test results can be found in [127]. In what follows we describe the objective measures we use in this thesis.

The aim of the objective measures is to provide a numerical value which allows to evaluate the merit of using one enhancement technique against another. That is, given a noisy input signal $x(n) = s(n) + w(n)$, where $s(n)$ is the clean speech signal and $w(n)$ is the noise signal, an estimate $\hat{s}(n)$ for $s(n)$ can be obtained from the noisy signal via some

filtering operation which can be represented as

$$\hat{s}(n) = \mathcal{H}_x \{x(n)\} \tag{8.1}$$

where $\mathcal{H}_x \{\cdot\}$ represents the series of filters derived from the noisy signal (hence the subscript x) on a frame by frame basis.

The most popular objective measure is the segmental SNR. In general, the SNR within one frame (with index t) is defined as the ratio of the speech energy to the noise energy. Hence, the (instantaneous) SNR in dB for the t^{th} frame would be given by

$$\mathrm{SNR}_t(s,w) = 10\log_{10}\sum_{n=0}^{L-1}|s(tD+n)|^2 - 10\log_{10}\sum_{n=0}^{L-1}|w(tD+n)|^2 \tag{8.2}$$

where L is the frame length and D is the frame shift usually chosen as $D = L/2$ (for a 50% overlap). The input segmental SNR, denoted as $\mathrm{SNR}(s,w)$, is then obtained by taking the arithmetic mean of the individual local SNR's per frame, in dB, over all available frames in the given speech signal[1] , i.e.

$$\mathrm{SNR}(s,w) = \frac{1}{T}\sum_{t=0}^{T-1}\mathrm{SNR}_t(s,w), \quad \text{in dB} \tag{8.3}$$

where T is the number of speech activity frames in the signal under consideration.

The output segmental SNR is calculated as $\mathrm{SNR}(s, s-\hat{s})$ which is the energy ratio of the clean speech to the residual error signal which includes residual noise as well as signal distortion. The noise reduction performance and the signal distortion can alternatively be measured using two separate measures by proceeding in the following way.

When the filtering process is linear (as is the case with the methods considered in this thesis) then $\mathcal{H}_x\{\cdot\}$ in (8.1) can be written as

$$\hat{s}(n) = \underbrace{\mathcal{H}_x \{s(n)\}}_{\hat{s}_r(n)} + \underbrace{\mathcal{H}_x \{w(n)\}}_{\hat{w}_r(n)} \tag{8.4}$$

where the output signal components $\hat{s}_r(n)$ and $\hat{w}_r(n)$ correspond to the input $s(n)$ and $w(n)$

[1]The segmental SNR is then the geometric mean of the energy ratios before transforming them into the log domain.

respectively. During simulations, the noise $w(n)$ is artificially added to the clean signal $s(n)$ to form the noisy signal $x(n)$. This latter signal is used to calculate the suppression filters represented by the operation $\mathcal{H}_x\{\cdot\}$. The filtering itself, however, is performed on every component separately, that is

$$\hat{s}_r(n) = \mathcal{H}_x\{s(n)\} \tag{8.5}$$

$$\hat{w}_r(n) = \mathcal{H}_x\{w(n)\} \tag{8.6}$$

Doing so, the linearity property of the filters is used to obtain the total output signal as $\hat{s}(n) = \hat{s}_r(n) + \hat{w}_r(n)$. With this scheme, the noise reduction capabilities and the incurred signal distortion can be measured separately as described next.

The noise reduction capabilities of one method is measured using the noise reduction factor (NRF) defined as

$$\text{NRF} = \text{SNR}(w, \hat{w}_r) \tag{8.7}$$

That is, it is the energy ratio of the input noise signal to the output residual noise signal. The signal distortion (DIST) is measured using the cepstral distance as follows

$$\text{DIST} = \text{cepd}(s, \hat{s}_r) \tag{8.8}$$

where the cepstral distance is defined as

$$\text{cepd}(s, \hat{s}_r) = \sum_{j=1}^{2J} (c(j) - \hat{c}(j))^2 \tag{8.9}$$

where $c(j)$ and $\hat{c}(j)$ are the cepstral coefficients of $s(n)$ and $\hat{s}_r(n)$ respectively in one frame and J is the model order chosen to be equal to 8 [105]. The cepstral coefficients are calculated according to the method described in [127].

In the multi-microphone case the output signal components are obtained as

$$\hat{s}_r(n) = \mathcal{H}_x\{s_1(n), \ldots, s_M(n)\} \tag{8.10}$$

$$\hat{w}_r(n) = \mathcal{H}_x\{w_1(n), \ldots, w_M(n)\} \tag{8.11}$$

where M is the number of microphones. In this context the NRF is measured as $\text{SNR}(w_1, \hat{w}_r)$

and the signal distortion as $\text{cepd}(s_1, \hat{s}_r)$. In the case of reverberated signals, we still use $s_1(n)$ as a reference to measure the signal distortion since the dereverberation problem is not covered in this thesis.

The method with the best performance will have the highest NRF while maintaining the minimal signal distortion. Note that there are available some standardized techniques to measure the quality of speech signals such as the PESQ approach by ITU [75]. However, to our knowledge, this standard, originally developed to evaluate speech coding methods, has not been validated for speech enhancement evaluations. For this reason we chose not to use this measure in our experiments.

8.2 Experimental Setup

8.2.1 Test sentences

For performance evaluation we used female and male sentences the content of which are shown in Table 8.1. These sentences will often be referred to by the acronyms given in the table. Sentences S1 and S2, used in one of the tests described, were recorded by both female and male speakers. The other sentences which are used more often in the experiments, are only recorded by female speakers for F1-3 and male speakers for M1-3. All the recordings were sampled at $F_s = 8$ KHz.

Table 8.1 Sentences used for performance evaluation.

Female sentences:		
F1	:	*Cats and dogs each hate the other.*
F2	:	*A lathe is a big tool.*
F3	:	*Grab every dish of sugar.*
Male sentences:		
M1	:	*Post no bills on this office wall.*
M2	:	*Primitive tribes have an off beat attitude.*
M3	:	*Live wires should be kept covered.*
Male & Female sentences:		
S1	:	*The ship was torn apart on the sharp reef.*
S2	:	*Sickness kept him home the third week.*

Unless otherwise mentioned, the results reported (specifically those involving objective measures) are based on the average of the values corresponding to each of the six sentences F1-3 and M1-3. During our research, other sentences were also used. However, we found that increasing the number of sentences does not significantly alter the results and it merely smoothes the obtained curves without changing the conclusions that can be drawn from them. Therefore, to save the simulation time, we decided to restrict our reported experimental results to these six phonetically rich sentences.

Since voice activity detection is beyond the scope of this thesis, the start and end points of speech for all these sentences have been manually labeled. Through out the experiments, the noise estimate is obtained using signal samples before the start of speech. Evaluation has been performed after clipping the non speech activity periods.

8.2.2 Noise types

The performance of the tested methods was evaluated under different noise types and SNR levels. The noise types used are presented in Table 8.2. For simplicity, these noises will sometimes be referred to with the corresponding acronyms shown in the table.

Table 8.2 Noise types used for performance evaluation.

WHT	:	Computer generated white noise.
VLV	:	Volvo car noise.
LEO	:	Leopard military vehicle noise.
JET	:	F16 jet cockpit noise.
FRZ	:	Freezer motor noise.
CMP	:	Computer fan noise.
KCH	:	Kitchen fan noise.
DRY	:	Dryer noise.

Again, throughout the research, other noise types were tested with similar observed results. However we restrict our reported experiments to the noises in Table 8.2 because they were found to be representative of the general behaviour of the tested methods.

While evaluating PSS and FBEVD, we have used WHT, JET, LEO, VLV and FRZ noises. The Blackman-Tukey PSD estimates of the last four (colored) noises are shown in Figure 8.1. The evaluation of the MEDA method was performed with the WHT, FRZ, CMP, KCH and DRY noises. The reason behind this choice is that in the multi-microphone

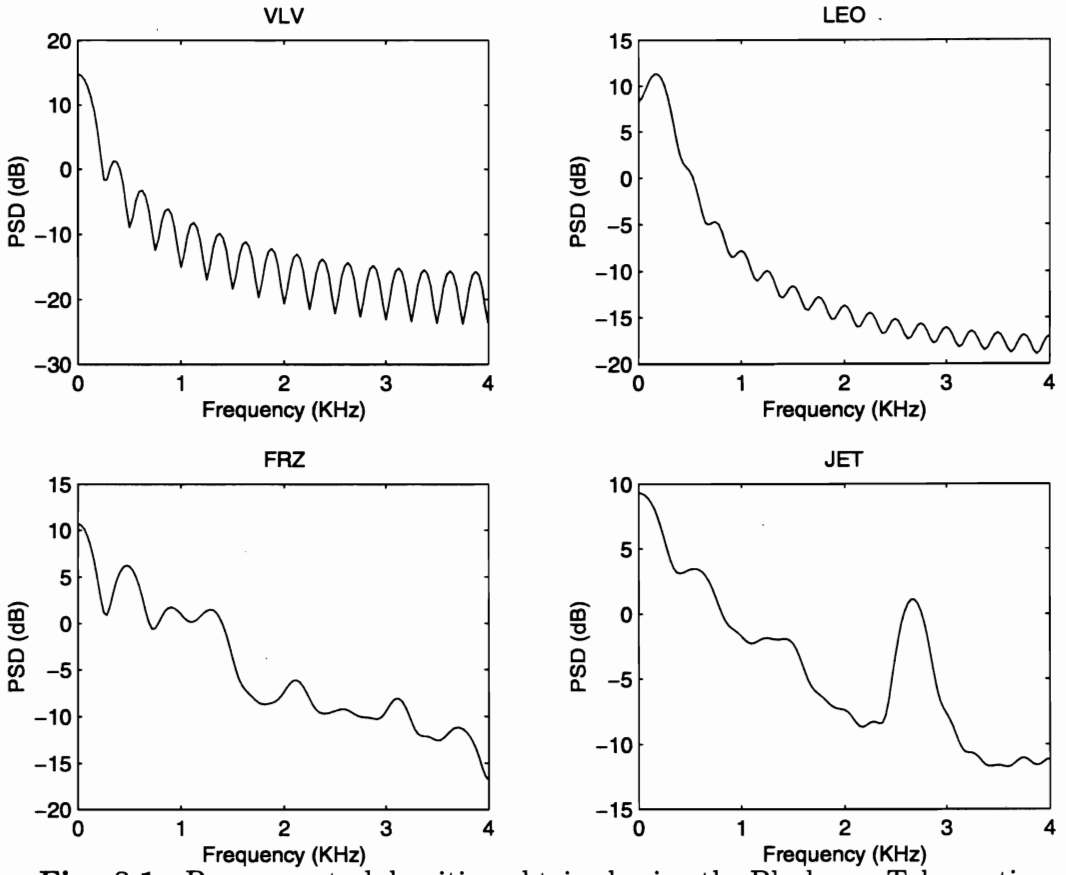


Fig. 8.1 Power spectral densities, obtained using the Blackman-Tukey estimator, of four colored noise types.

case, a room environment is tested. In such a setup, it would not be realistic to have a car engine noise, for instance, as the noise source.

8.2.3 Parameter values

- Frame length: $L = 256$

  At 8 KHz, the sampling rate used in all experiments, the frame length would be 32 msec which is good enough for the assumed speech stationarity.

- Model order: $P = 32$

  Our experiments showed that this value offers a good trade off between complexity and performance. Actually increasing P beyond this value does not significantly

improve the performance in a manner worth the added complexity. For comparison purposes, the same value is also used for MSVD[2].

- Synthesis window: Hanning

  Both Hanning and Hamming windows were tested. It was found that the former gives a slightly better performance. The difference, however, was not that much significant.

- Gain function:

  In general, the decaying exponential gain function (3.73) is found to have better noise reduction capabilities. This gain function is used in the single microphone case.

  However, for comparison purposes, we use the Wiener like gain function (6.57) in the multi-microphone case as stated in Section 6.6. This is because it resembles the gain functions used in the MSVD and the frequency domain Wiener adaptive postfiltering methods. The value of the control parameters for each method are specified later.

- DFT size: $K = 256$

  Experimental results showed that using a DFT with size 256 results in a better performance due mainly to a more accurate masking threshold. A detailed discussion on the value of K is given in Section 4.5.1.

8.2.4 Reverberation simulation

In the multi-microphone case, most experiments involve evaluation under reverberant conditions. The reverberated signals are obtained by the new subband room simulator (SRS) presented in Chapter 7. The room configuration used is shown in Figure 8.2. A linear array with $M = 4$ omni-directional microphones is used with a 5 cm inter-microphone distance. The microphones are assumed to have a flat frequency response equal to one. The microphones as well as the noise and speech sources are placed at a height of 1m. The room height is 3m. Unless otherwise mentioned, the speech and noise direction of arrival (DOA) are 0° and 45° respectively.

The speech and noise sources are placed at a distance far enough from the microphone array so that the far-field assumption is not violated. The far-field assumptions implies

[2]In fact the value of 20 is proposed in [35] where it is also reported that the performance actually improves by increasing the value of P (for example to 32).

that the signal attenuation due to propagation is constant over all microphones. For the used room configuration, and at 5 cm inter-microphone distance, the minimum distance between the sound source and the array would be according to (6.1)

$$r > \frac{F_s d_{M1}^2}{c} = \frac{8000(0.2)^2}{340} = 0.94 \text{ meters} \tag{8.12}$$

The selected inter-microphone distance will also guarantee that all frequencies above 340 Hz will be incoherent as given by (6.8).

The reflection coefficients of the walls, roof and floor are chosen to have the same value. This value will be changed according to the experiments to yield a different reverberation time. The reverberation, denoted as T_{60}, will be used to describe the amount of reverberation. The reverberation time is defined as the time required for the sound pressure level to decay to -60 dB of its original value. It can be measured for example using Sabine's formula [103]

$$T_{60} = \frac{0.163V}{S(1-\beta^2)} \tag{8.13}$$

where V is the room volume and S is the total surface area and β is the reflection coefficient, assumed to be constant for all the surfaces.

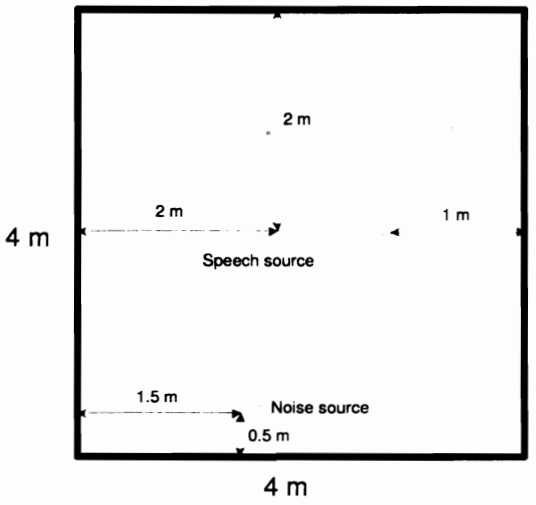


Fig. 8.2 Reverberant room setup.

In one particular case, frequency dependent reverberation times are assumed. Simulating such a scenario is made possible thanks to the novel SRS developed in Chapter 7.

The reverberation times per frequency band, shown in Table 8.2.4, are chosen so that they imitate the room acoustics described in [117]. This room will be referred to as the FDRT room, for Frequency Dependent Reverberation Time.

Table 8.3 The frequency dependent reverberation times, and the corresponding reflection coefficients, used in the FDRT room.

Frequency band (Hz)	Reverberation time (msec)	Reflection coeff.
0-250	650	0.923
250-1000	490	0.896
1000-2000	330	0.842
2000-4000	250	0.785

In the multi-microphone experiments all results are obtained for 5 noise realizations, 6 speech sentences, and are repeated for 5 noise types. Hence for every signal source (speech and noise) 5×5×6=150 filtering operations are performed for every tested reverberation time T_{60} or input SNR value to obtain the corresponding reverberant signal. For example, for 8 input SNR values, this filtering is repeated 150×2×8=2400 times in just one experiment. This large number revealed the significant benefit of using the novel SRS in reducing the computational load and allowing to perform these experiments in a considerably shorter period.

8.3 The Frame Based EVD method

In this section we evaluate the performance of the new Frame Based EVD implementation method (FBEVD) introduced in Section 5.2. In the signal subspace approach, overlapping vectors of length $P = 32$ are enhanced. In the original SSA implementation in [41], and as described in Section 3.4.4, a new signal subspace filter is calculated for every such vector implying that an EVD is carried out every $P/2 = 16$ samples resulting in a high computational cost. In the FBEVD scheme, it is assumed that the covariance matrix of the speech signal, hence its EVD, is relatively constant within a frame of length $L = 256$. Consequently, the same signal subspace filter can be used to enhance all the signal vectors within that frame. Since the frames have a 50% overlap, the EVD is only computed every

128 samples hence giving rise to considerable computational savings. The enhanced vectors are multiplied by a Hanning window and overlap-added to form the enhanced frame. The so obtained frames are then multiplied by a second Hanning window and synthesized with the overlap-add technique to yield the total enhanced speech signal.

8.3.1 Performance evaluation

To evaluate the performance of FBEVD, we run experiments using RQSS, described in Section 3.4.5, on the six sentences (F1-3, M1-3) and the five noise types (JET, LEO, VLV, FRZ, and WHT). The performance is measured using the noise reduction factor and the cepstral distance as explained in Section 8.1.

The experiment consists of testing the performance of RQSS as a function of the frame length L. The value varies from $L = P = 32$, as it is the case with the original SSA implementation, to $L = 256$, the frame length suggested in this thesis. To calculate the autocorrelation function, though, a window with a fixed number of samples (256 samples) is used in all cases. For FBEVD, this is accomplished by adding samples on both sides of the current frame to acquire the necessary number of samples. Note that when $L = P$, just a single synthesis Hanning window is used. That is, the Hanning window normally used during the frame synthesis is replaced by a rectangular window so that the same implementation as in the original SSA is retained.

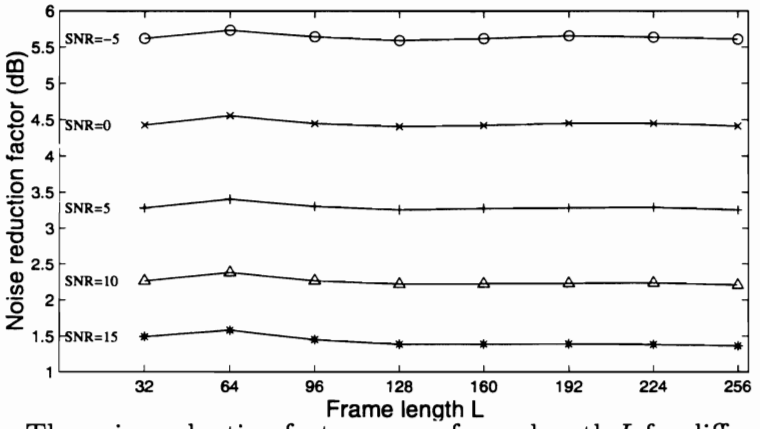


Fig. 8.3 The noise reduction factor versus frame length L for different input noise levels.

Figures 8.3 and 8.4 show the noise reduction factor and the signal distortion of the FBEVD for different input noise levels as a function of the frame length L. It can be

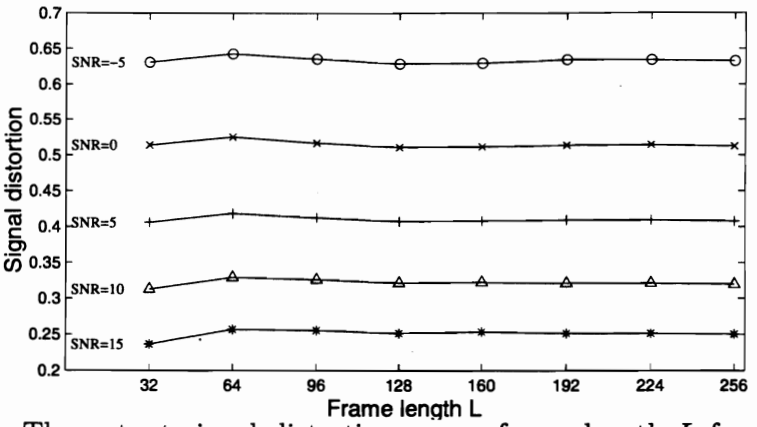


Fig. 8.4 The output signal distortion versus frame length L for different input noise levels.

seen that there is essentially no significant effect on the performance as the frame length increases. The particular cases of $L = 32$ and $L = 256$ (the original SSA implementation and the frame length used in this thesis, respectively) are shown in Figure 8.5. Again it can be seen that the FBEVD has almost no effect on performance. It should also be noted that, at low SNR conditions, informal listening tests reveal that increasing the frame size actually smoothes the residual noise making its musical character less annoying and the overall signal more pleasant to the human ear.

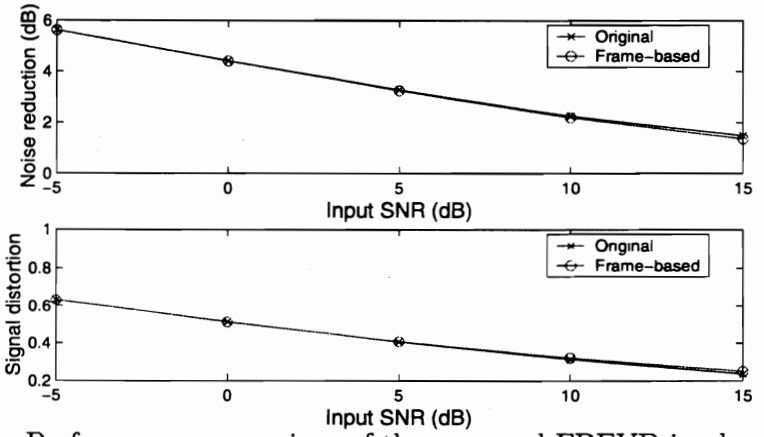


Fig. 8.5 Performance comparison of the proposed FBEVD implementation and the original SSA implementation in terms of noise reduction factor (up) and signal distortion (down).

8.3.2 Computational savings

We now try to quantify the computational savings achieved by the FBEVD method. We do that by measuring the time spent to process one speech file and divide the result by the total number of samples. Hence the quantity we use to compare the computational load for different frame lengths is the processing time per input sample (TPIS) in msec. These experiments were run on a 2.4 GHz Pentium IV processor with 512 MB of memory.

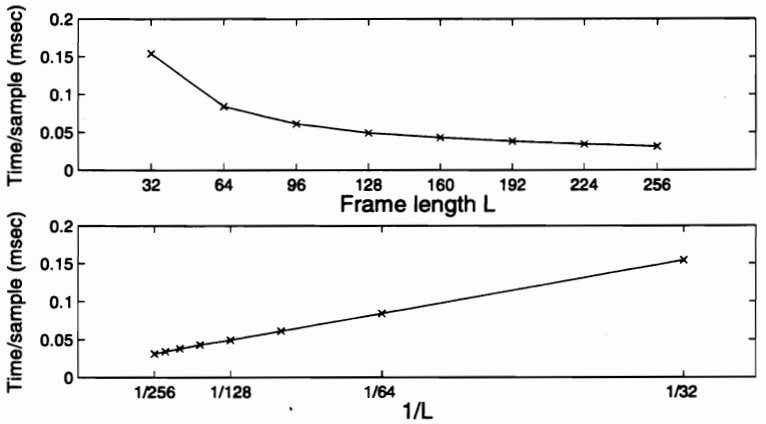


Fig. 8.6 The time per input sample needed by RQSS versus the frame length L (up) and versus $1/L$ (down).

It can be seen from Figure 8.6 that as the frame length increases the TPIS decreases, i.e. it is inversely proportional to the frame length L. For example, with the original SSA implementation ($L = P = 32$) about 0.16 msec per sample were needed whereas for $L = 256$ this figure goes down to 0.03 msec. Actually, from the bottom plot of Figure 8.6 it can be verified that the TIPS increases linearly with $1/L$. Therefore we conclude that compared to the original implementation, the computational savings achieved by FBEVD are proportional to L/P. Indeed this is expected since the frames have an overlap of 50% and there are $2L/P - 1$ vectors per frame.

Note also that at 8 KHz sampling rate, the sample period is about 0.12 msec so we can say that even with a Matlab implementation, and thanks to the available processing power, the proposed method was able to bring the required computational burden to an affordable real time level.

In Figures 8.7 and 8.8 we again show the required TPIS versus L and $1/L$ respectively but this time we seek to observe the added computational load due to the use of masking. The curves show TPIS for RQSS as well as PSS with three different masking models,

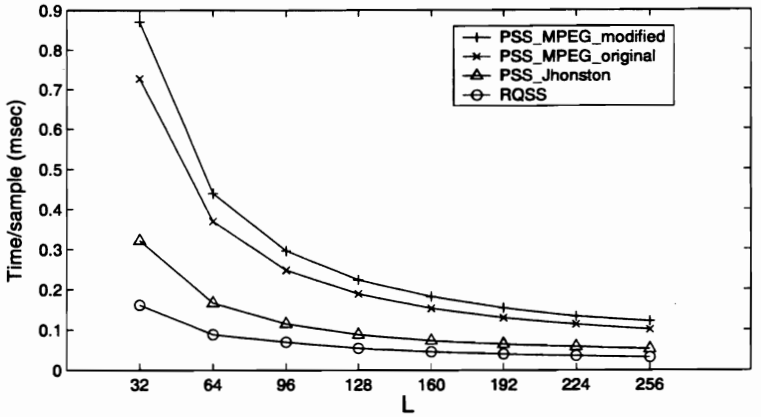


Fig. 8.7 The time per input sample needed by RQSS and PSS (with different masking models) versus the frame length L.

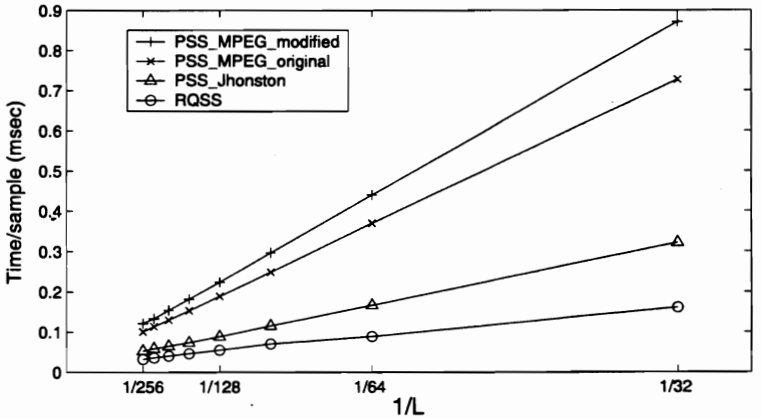


Fig. 8.8 The time per input sample needed by RQSS and PSS (with different masking models) versus $1/L$.

namely, Jhonston's model, the original MPEG model and the modified MPEG model[3]. It can still be concluded that in all four situations the required processing time is increasing linearly with $1/L$.

Moreover, we can conclude from these figures that with the FBEVD method, the added complexity due to the masking threshold calculation and all the subsequent calculations such as computing the matrix $\mathbf{V}$ via FFT[4] is negligible. For $L = P$ however, the added complexity is very high which may render the proposed PSS method not suitable in practice

[3]These models are presented in Chapter 2 and Section 5.4.

[4]The matrix $\mathbf{V}$, needed for FET and defined in (4.25), has on its columns the magnitude squared DFT of the eigenvectors of the clean covariance matrix $\mathbf{R}_s$.

if FBEVD was not used (see $L = 32$ in Figure 8.7). Besides, we note that the modifications we made on the MPEG model, as discussed in Section 5.4, did not result in serious complexity increase despite the high frequency resolution used as compared to the one bark resolution of the original model.

Finally we note that the proposed PSS method with the FBEVD implementation, in addition to the improved speech enhancement performance it offers (as we will see shortly), is actually faster than RQSS when implemented with the original SSA scheme, i.e $L = P$. These computational savings can be further improved by employing a fast EVD method and also by using a more efficient masking model.

8.4 The Perceptual Signal Subspace method

In this section we evaluate the gain in performance achieved by the novel Perceptual Signal Subspace method (PSS) presented in Chapter 5. The evaluation is based on informal listening tests and spectrogram illustrations and especially on different subjective tests. Since a fundamental concept in the proposed method is that as long as the corrupting noise is not audible, it is allowed to stay in the enhanced signal, we found it more instructive to base our evaluation on subjective tests rather than objective measures. These subjective tests are mainly A-B tests where the listener has to choose a preferred sentence from pairs of recordings representing different enhancement methods. Another subjective test measuring a quantity called the "noise shaping score" is designed in this thesis and is used to measure the capability of the different methods to shape the spectrum of the residual noise according to the desired speech signal spectrum.

PSS is evaluated against two other methods. Namely the Raleigh Quotient method presented in Section 3.4.5 in which the noise energy along every eigen-direction is used to handle the colored noise case. This approach is the basis for the methods described for example in [130] and [121]. The second method the PSS is evaluated against is the original SSA method with prewhitening (PWSS) presented in [41] and described in Section 3.4.5 of this thesis.

In order to maintain a relatively constant signal distortion across the different methods, the control parameter in the decaying exponential gain function (3.73) was set to $\nu = 0.8$ in PSS and $\nu = 1$ in RQSS and PWSS. This choice is also made so that, for comparison purposes, some residual noise can still be audible.

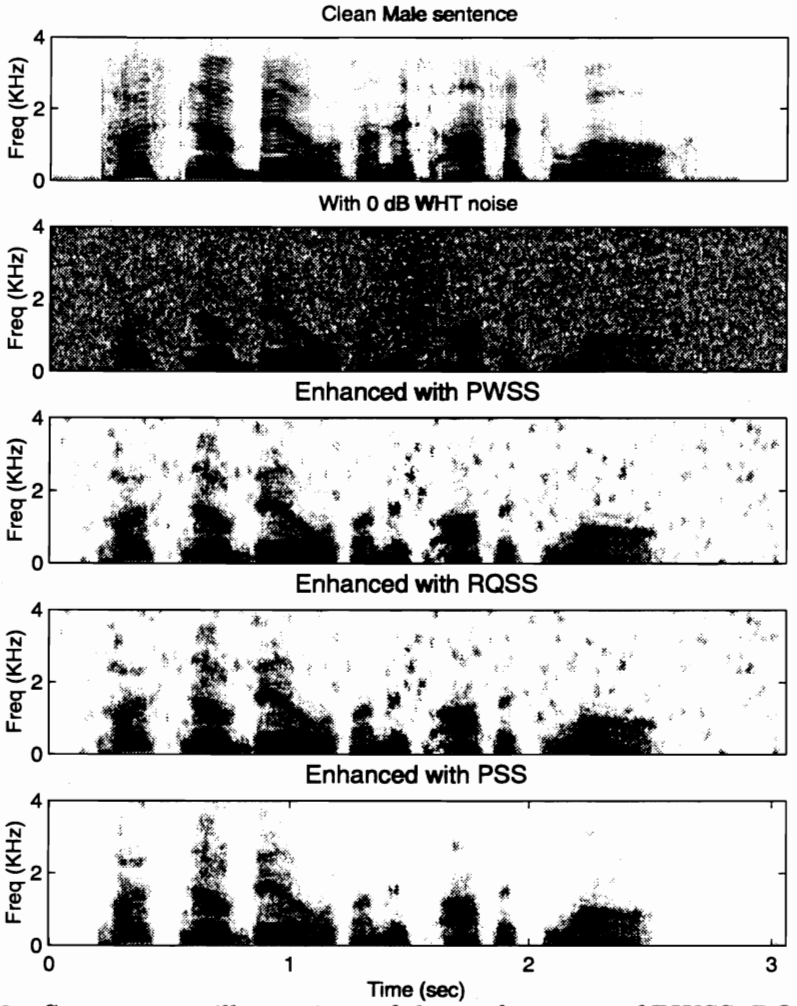


Fig. 8.9 Spectrogram illustrations of the performance of PWSS, RQSS and PSS on the Male sentence when corrupted with white noise.

8.4.1 Informal listening tests and spectrogram

During informal listening tests, the performance of PSS was evaluated against PWSS and RQSS on different sentences corrupted with several noise types. This evaluation revealed the superiority of our proposed method. Indeed, PSS resulted in a less audible residual noise while maintaining a similar level of signal distortion as the competing methods.

This claim is supported by the spectrogram illustrations shown in Figures 8.9, 8.10, 8.11 and 8.12 for WHT, FRZ, JET and LEO noises respectively. The sentence shown is the male sentence M1. It can be seen from these figures that all the formants, important for intelligibility, are preserved while the noise is almost completely canceled whenever the

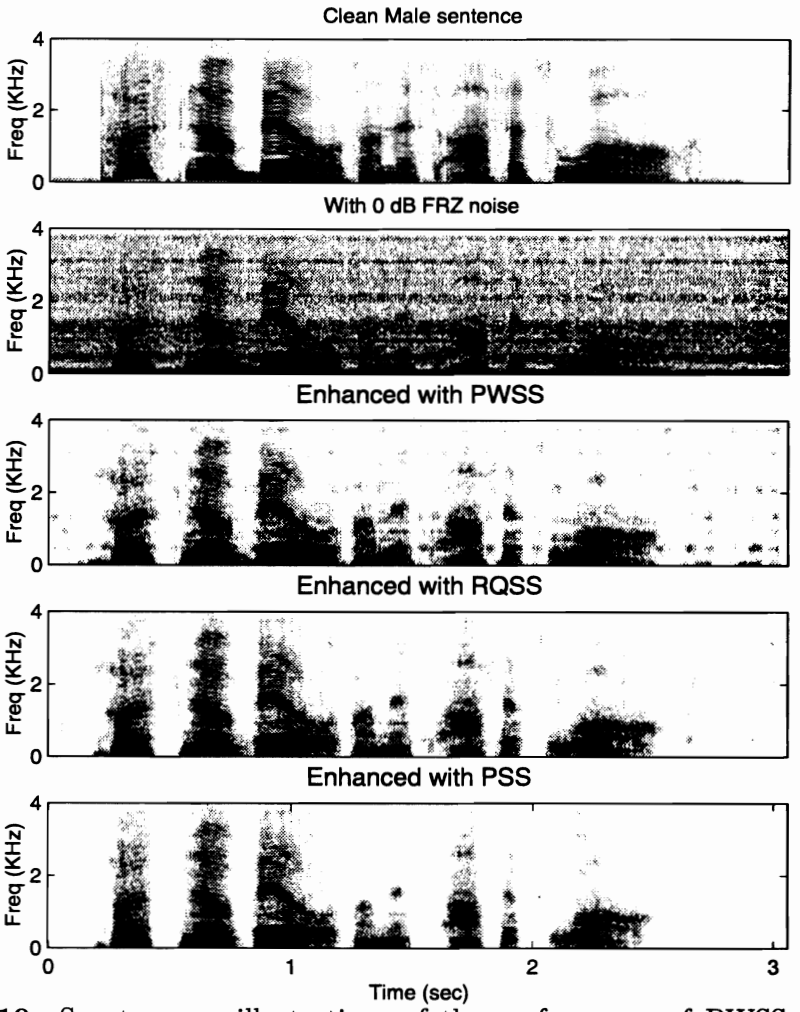


Fig. 8.10 Spectrogram illustrations of the performance of PWSS, RQSS and PSS on the Male sentence when corrupted with freezer motor noise.

speech is absent. For PWSS and RQSS, on the other hand, the musical noise is clearly seen (and indeed heard) especially for white noise where the random tones are present in all frequency bands due to the flat spectrum of the original corrupting noise.

In the case of JET noise for example, shown in Figure 8.11, the high frequency peak (around 2.8 KHz) as can be seen in Figure 8.1, is still present for PWSS and RQSS which almost completely failed to cancel the noise at that frequency. In fact, even for PSS, that peak seemed to reduce the benefit of using masking as compared to other noise types, and the achieved improvement was not that much evident. This might be explained by the fact that the high frequency peak might have affected the accuracy of the masking threshold

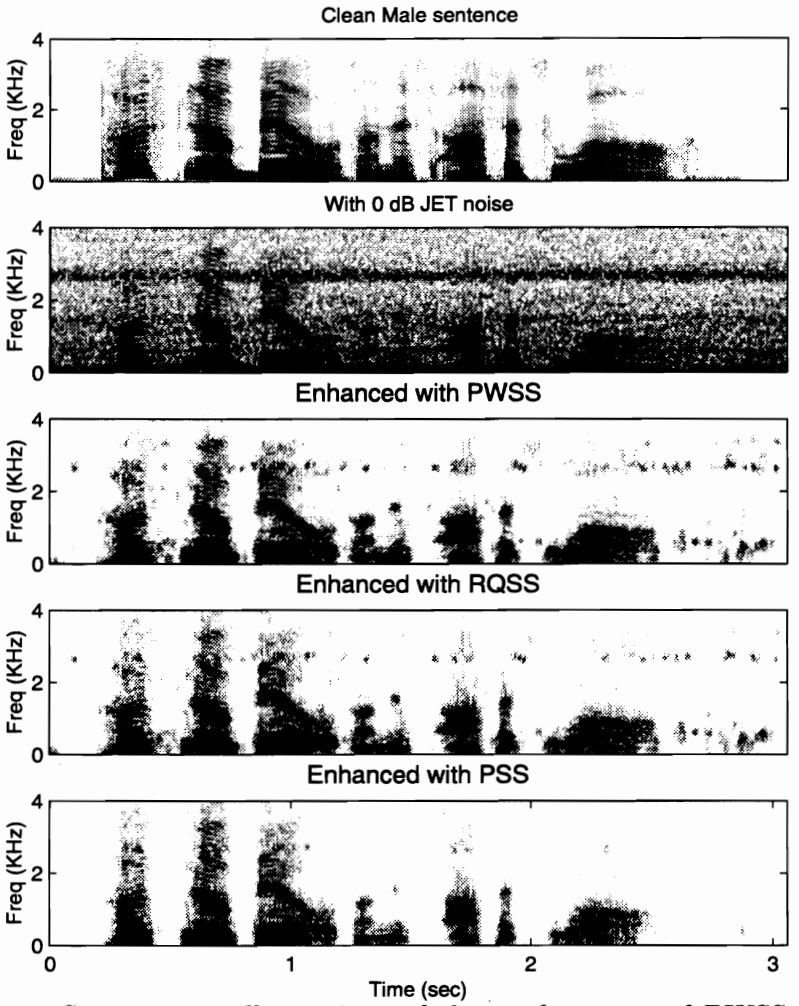


Fig. 8.11 Spectrogram illustrations of the performance of PWSS, RQSS and PSS on the Male sentence when corrupted with F16 jet cockpit noise.

estimate, resulting in a lower performance compared to other noises.

Another important result observed during informal listening tests, is that with PSS, the perceived residual noise has relatively the same spectral characteristics whatever was the original corrupting noise. Note that we deliberately chose the value of the control parameter ν in the gain function so that the signal distortion be at an acceptable level and be more or less the same across the three methods while allowing the residual noise to be audible to some extent (that is, not completely suppressed).

For PSS the residual noise kept little of its original character with its energy increasing and decreasing according to the energy of the speech signal itself. Actually for this reason,

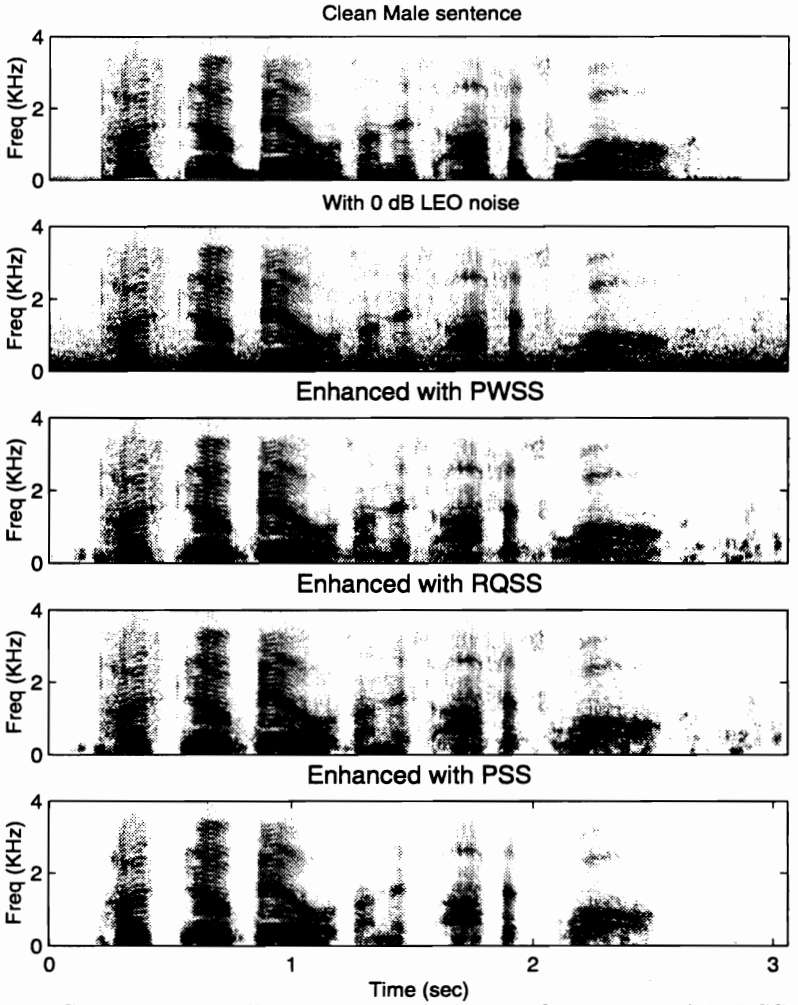


Fig. 8.12 Spectrogram illustrations of the performance of PWSS, RQSS and PSS on the Male sentence when corrupted with Leopard vehicle noise.

as explained in Section 5.3, equation (5.7), we used the minimum operator in an attempt to make the transition between speech activity and non-activity periods smoother and hence more natural, resulting in a more pleasant signal.

The spectrogram illustrations back this claim where "visually" one cannot distinguish signals corrupted by different noises after enhancing it with PSS. This, again according to the spectrograms, is not true for PWSS and RQSS where the difference can be easily seen. We refer to this phenomenon as *noise shaping* which describes the ability of the enhancement method to give the spectrum of the residual noise a shape which resembles that of the speech signal itself, hence making the noise inaudible by masking it without

Table 8.4 A-B test 1: White noise at different input SNR levels. Shown are the percentage of times where PSS was preferred, compared to the noisy signal and the original SSA.

Input SNR	Compared with noisy signal	Compared with SSA
5 dB	92%	71%
-5 dB	85%	78%
-10 dB	85%	92%

being completely suppressed. The above mentioned result will be further supported by the noise shaping score obtained via a subjective test which we conceived to reveal this particular property (see Section 8.4.5).

We note also that PWSS and RQSS are very sensitive to the value of ν which controls the trade-off between the signal distortion and the level of the residual noise. The best performance is achieved by tuning that parameter according to the application conditions, that is the noise type, the SNR level and the speech utterance itself. For PSS, on the other hand, such tuning is not necessary and the masking threshold seems to take care of it. Actually ν is chosen to be small for PSS because the masking threshold is already lowered by subtracting the masking offset from it. This masking offset, as presented in Chapter 2, depends on the tonality of the signal, that is whether it is noise like or tone like. Therefore, the value of ν is indirectly tuned to suit the situation under consideration, which is exactly the aim of this approach.

8.4.2 A-B test 1: White noise

In this subjective test carried out during an early stage of our research, the performance of PSS on signals corrupted by white noise for different SNR levels, was evaluated[5]. In this experiment, PSS was implemented using Johnston's masking model.

The following methodology was used: Sentences S1 and S2 were spoken by the same speaker and played from a single file, separated by a short pause. Four such speech files obtained from two male and two female speakers were used. Each of the four original recordings was corrupted with additive computer generated white noise at three input SNR levels (5 dB, -5 dB and -10 dB). The 12 test files so obtained were enhanced using

[5]The results of this test have been reported in [79].

Table 8.5 A-B test 2: Colored noise case with four different noise types. Shown are the percentage of times where PSS was preferred, compared to noisy signal, PWSS and RQSS.

Noise Type	SNR (dB)	Compared with noisy signal	Compared with PWSS	Compared with RQSS
FRZ	-4	100%	90%	80%
LEO	-4	100%	100%	100%
VLV	-10	80%	85%	60%
JET	0	60%	70%	60%

the proposed PSS method and the SSA. The SSA here denotes the method proposed by Ephraim and Van Trees, that is PWSS but in this case, since the noise is already white, no prewhitening is required.

For every test file two comparisons were made: PSS versus SSA and PSS versus noisy speech, leading to a total of 24 pairs. 14 subjects participated in the test among which there were two who worked in the speech processing field but were not familiar with the sentences. The subjects were asked to compare the two recordings of every pair and choose the one they prefer.

Table 8.4 shows the results of this test. On the average the subjects voted for the proposed PSS method over the noisy signal 87% of the times and over the SSA 80% of the times. The PSS becomes more useful at very low SNR conditions where the subjects preferred the use of masking to enhance the speech signals 92% of the times.

8.4.3 A-B test 2: Colored noise I

A second subjective A-B test has been carried out to verify the performance of PSS under different types of colored noise. In this test, the 2.2 sec long F1 female sentence has been used[6].

The noises tested were FRZ, VLV, LEO and JET. The different input SNR levels for every noise type are shown in Table 8.5. Another group of 14 people[7] were asked to evaluate the performance of the new PSS method and to compare it with the original noisy signal, the PWSS and the RQSS. None of the subjects worked in the speech processing field. 12

[6]The results of this test have been reported in [80].

[7]The subjects are different from those who participated in A-B test 1.

pairs of recordings were presented to the subjects: for each pair, they were asked to vote for the signal they preferred. In this test a neutral answer was also allowed if they could not perceive any difference.

Table 8.5 shows the results of this test. It can be seen that PSS outperformed the other two enhancing methods especially with LEO noise were all the subjects voted for PSS. We note that in the JET noise case 40% of the subjects voted for the noisy signal because they preferred the existing noise to the obtained signal distortion. However, these subjects said that if the 2.2 sec test signal had been longer they would have changed their preference because the noise would be more disturbing and they would be less able to tolerate it.

8.4.4 A-B test 3: Colored noise II

In A-B test 2, the results showed the superiority of our proposed method over the competing methods. However, due to the test's design, the results do not reveal the criteria which were actually behind the achieved improvement. That is, whether the improvements were due to a lower signal distortion, a lower residual noise energy or both. For this reason, we decided to repeat the test and modify it so as to allow the subjects to tell what exactly they preferred in the enhanced signals.

In this test we again used sentence F1 (the Female sentence) and also added sentence M1 (the male sentence) which is 3 sec long. The same four noises were used, that is FRZ, VLV, LEO and JET, all at 0 dB segmental SNR except VLV which was at -5 dB. The number of people who took part in the test was 18 among which three worked in the speech processing area but were unfamiliar with the sentences. The majority of the subjects were in their late twenties.

In total, 8 pairs of recordings per test were presented to the subjects where each pair consisted of a speech signal enhanced using PSS and a second enhanced with a competing method, namely RQSS and PWSS . A separate test has been conducted for every sentence. For each pair, they were asked to vote for the signal they preferred (A, B or X if they had no preference) according to three different criteria:

- Intelligibility: *Which signal was easier to understand?*
- Quality : *Which signal was less noisy?*
- Overall: *Putting the previous two criteria together, which signal was preferred?*

Table 8.6 A-B test 3: preference results for the Female sentence.

Noise	Signal Distortion					
	PSS	X	PWSS	PSS	X	RQSS
FRZ	**20%**	73%	7%	**20%**	73%	7%
VLV	**33%**	40%	27%	**27%**	40%	33%
JET	**14%**	50%	36%	**33%**	47%	20%
LEO	**53%**	40%	7%	**33%**	53%	13%
Noise	Residual Noise					
	PSS	X	PWSS	PSS	X	RQSS
FRZ	**67%**	27%	7%	**60%**	40%	0%
VLV	**73%**	20%	7%	**47%**	40%	13%
JET	**57%**	43%	0%	**47%**	47%	7%
LEO	**87%**	0%	13%	**80%**	13%	7%
Noise	Overall					
	PSS	X	PWSS	PSS	X	RQSS
FRZ	**67%**	27%	7%	**60%**	33%	7%
VLV	**80%**	13%	7%	**67%**	20%	13%
JET	**50%**	43%	7%	**47%**	40%	13%
LEO	**87%**	13%	0%	**80%**	20%	0%

Tables 8.6 and 8.7 show the results of this test for the female and male sentences respectively. It can be seen that the PSS method outperforms the other two methods especially for the female sentence. In general, the subjects found that the three methods provided a relatively similar amount of distortion to the enhanced signals, with the exception on the Female-LEO and Male-VLV cases where the use of PSS resulted also in a less distorted signal than PWSS. Overall, the merit of PSS is in that it succeeds to maintain an acceptable level of distortion while offering a better noise reduction (masking) performance.

Nonetheless, it should be noted that some of the subjects voted in favor of PSS because they find that the corresponding enhanced signals had less residual noise and also was "easier to understand". This is because the presence of noises made it more difficult to figure out what has been said. That is, according to these subjects, intelligibility and quality are two very related features.

PSS had a considerable success over RQSS and PWSS in the case of LEO, VLV and to a less extent FRZ. For example in the LEO case, the subjects preferred PSS over RQSS 80% and 67% of the times for the female and male sentences respectively. Compared with

Table 8.7 A-B test 3: preference results for the Male sentence.

Noise	Signal Distortion					
	PSS	X	PWSS	PSS	X	RQSS
FRZ	**28%**	56%	17%	**28%**	61%	11%
VLV	**72%**	22%	6%	**47%**	29%	24%
JET	**17%**	61%	22%	**11%**	72%	17%
LEO	**39%**	39%	22%	**50%**	28%	22%
Noise	Residual Noise					
	PSS	X	PWSS	PSS	X	RQSS
FRZ	**67%**	22%	11%	**50%**	33%	17%
VLV	**89%**	11%	0%	**59%**	41%	0%
JET	**44%**	39%	17%	**28%**	67%	6%
LEO	**89%**	6%	6%	**83%**	17%	0%
Noise	Overall					
	PSS	X	PWSS	PSS	X	RQSS
FRZ	**61%**	33%	6%	**56%**	39%	6%
VLV	**89%**	6%	6%	**71%**	18%	12%
JET	**39%**	44%	17%	**39%**	56%	6%
LEO	**83%**	11%	6%	**67%**	17%	17%

PWSS, these figures were at 87% and 83% respectively. Note that when the subjects *did not* vote for PSS, that was mostly because they were unable to perceive any difference rather than because PSS had a poorer performance.

In the JET case, the improvement achieved over the two competing methods was not as obvious as it is with the other noises. The test results revealed that many subjects were not able to perceive such improvement especially compared with RQSS. In fact, due to the characteristics of the JET noise (and to its high level), the masking threshold estimate was poor resulting in inaccurate values of the perceptual energies θ_i's which turned out to be most of the time larger than the original eigenvalues λ_i's. Since the gain function (5.7) takes the minimum of the two, the masking threshold becomes ineffective and PSS behaves in a close manner to RQSS. If the minimum is not utilized, serious clipping of important speech parts occurs leading to an undesired signal distortion. Hence, the minimum operator acts as a protector that ensures that PSS would at least have the same performance as RQSS or PWSS whenever the masking threshold estimate is not very accurate.

8.4.5 The residual noise shaping score

As mentioned earlier, during informal listening tests, we have observed that signals enhanced with PSS have residual noise characteristics which are relatively similar, regardless of the original corrupting noise. This result supports our claim that PSS yields improved noise shaping and hence better masking. Noise shaping here means that the residual noise spectral characteristics have been modified, or shaped, by the enhancement method in a way that perceptually it sounds as close as possible to the desired speech signal. Consequently, a speech signal corrupted by two different noises would, after enhancement, arguably have the residual noises in both cases sound relatively the same.

To confirm this result, we have conceived a new subjective test which provides a "*residual noise shaping score*" serving to compare the performances of the different methods according to the above mentioned criterion.

The subjects were presented with a pair of signals enhanced by the *same* method but corresponding to different noises. Then they were asked to *concentrate* just on the background noise and to compare its characteristics in the two recordings. The comparison is based on how similar or different these characteristics are in the two signals, regardless of the loudness. The subjects had to score their decision according to a 5-level rating scheme as shown in Table 8.8. Again we have used the same four noises resulting in 6 pairs for every method. In total, for the three methods, 18 pairs per test were presented to the subjects. Two tests, one for every sentence, had been designed.

Due to this relatively high number, the subjects were asked to make their decisions after listening to a group of three pairs at a time. Every group corresponded to a different noise pair and every pair within one group corresponded to one of the three enhancing methods[8]. The aim of doing so was to help the subjects establish a kind of reference. This makes the scores a relative measure rather than an absolute one.

Table 8.8 Rating scheme for the residual noise shaping score test.

1	:	Completely different
2	:	Different
3	:	Don't know
4	:	Similar
5	:	Very similar

[8]This grouping criterion was of course not revealed to the subjects.

Table 8.9 Residual noise shaping scores for the female (F) and male (M) sentences.

Noise pair	PSS		RQSS		PWSS	
	F	M	F	M	F	M
LEO, FRZ	4.3	4.2	2.3	2.2	1.5	1.1
VLV, FRZ	4.1	3.8	2.5	2.1	1.2	1.4
LEO, VLV	4.7	4.5	2.1	1.8	1.5	1.7
LEO, JET	3.3	2.8	1.5	1.2	1.8	1.6
VLV, JET	3.0	2.7	2.1	1.9	2.0	2.1
JET, FRZ	3.8	3.7	3.1	2.9	2.7	3.1
Average	**3.9**	**3.6**	**2.3**	**2.0**	**1.7**	**1.8**

The detailed scores for the different noise pairs for the two sentences are given in Table 8.9. It can be seen that PSS got a higher score on average than RQSS and PWSS which shows that it achieves a relatively better noise shaping than the other two competing methods.

8.5 The Multi-microphone Eigen-Domain Averaging method

In this section, we evaluate the performance of the new proposed Multi-microphone signal subspace method with Eigen-Domain Averaging (MEDA) under different conditions. The MEDA is compared against a competing signal subspace method, namely the MSVD method from [35, 34] which was described in Section 6.4.4. In order to verify the benefit of using eigendomain averaging, MEDA is also compared against the MRQSS presented in Section 6.5 and which constitutes the basis of the MEDA method. This comparison allows to assess the merit of the eigendomain averaging technique which constitutes the main difference between MEDA and MRQSS .

Another approach we evaluate MEDA against is when RQSS is applied individually to every channel and the average output of the M filters is taken as the estimated speech signal[9]. This approach has been suggested in [61] as a generalization of the SSA into a multi-microphone design. We refer to this method as SSM where M is the number of microphones. For instance SS4 is RQSS applied to 4 microphones, the number used in the

[9]RQSS is the single channel signal subspace method which uses the Raleigh Quotient approach to handle the colored noise case.

experiments. More accurately, the SSM estimate is obtained as follows

$$\hat{\mathbf{s}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{H}_m \mathbf{x}_m \tag{8.14}$$

where $\mathbf{H}_m$ is the RQSS filter calculated in the m^{th} channel.

As will be demonstrated shortly, MEDA is particularly useful under diffuse noise conditions. For this reason, its performance is also compared to that of the popular frequency domain adaptive Wiener postfilter described in Section 6.4.3. This method, referred to here as the Wiener method, is mainly based on the algorithm presented in [170] and adopted in many variant methods including [117]. As discussed in Section 6.4.3 this method is known to have a good performance in diffuse noise fields.

As stated in Section 6.6 the Wiener like gain function 3.72 is used in MRQSS, MEDA and SS4. The control parameter is chosen to be $\mu = 0.8$ for MEDA and MRQSS and $\mu = 0.6$ for SS4. This choice is made in order to obtain, whenever possible, the same signal distortion across all methods. The evaluation is then generally based on the noise reduction factor (NRF). Refer to Section 8.1 for definition of the signal distortion and the noise reduction factor used here.

In all experiments the results are given for white noise and colored noise. The latter case is obtained as the average over 4 noise types, namely FRZ, DRY, KCH and CMP (see Table 8.2). The individual detailed performance of these noises is given in Appendix B.

Note that MEDA, MRQSS and SS4 are all implemented using the fast FBEVD implementation scheme developed in this thesis and described in Section 5.2. MSVD is implemented using the batch (non-recursive) mode described in [34].

8.5.1 Performance versus input SNR level

To evaluate the performance of MEDA as a function of input SNR, three experiments were conducted. The three experiments were carried out using the room configuration described in Section 8.2.4 and depicted in Figure 8.2. In every experiment the reverberation conditions are changed in order to test a different environment.

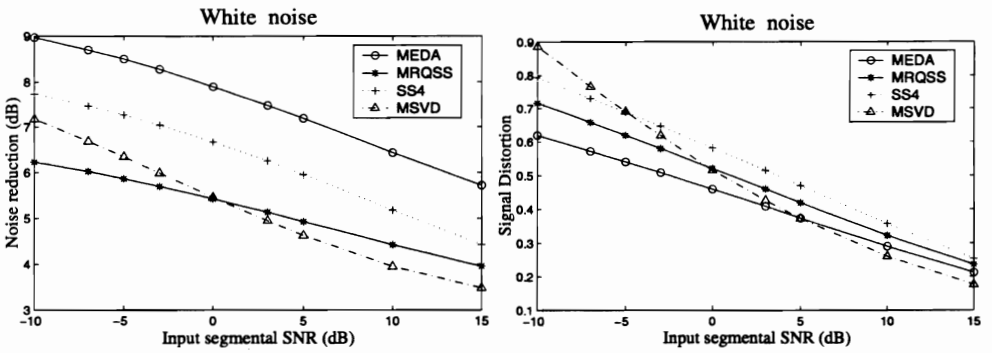


Fig. 8.13 Performance evaluation under different input segmental SNR levels at 400 msec reverberation time for white noise.

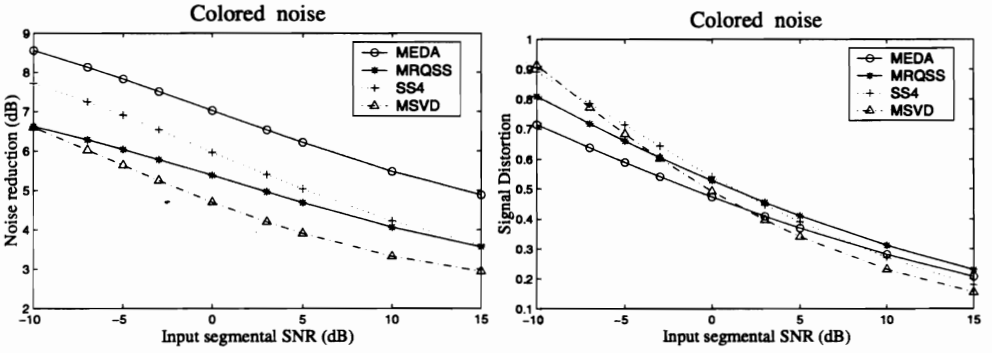


Fig. 8.14 Performance evaluation under different input segmental SNR levels at 400 msec reverberation time for colored noise.

Experiment 1

In the first experiment, the performance under relatively reverberant conditions is assessed. The reverberation time is set to $T_{60} = 400$ msec which can be considered to be within the range of a typical room. Figure 8.13 shows the results for white noise while Figure 8.14 shows the results for colored noises. In this experiment, MEDA is evaluated against MSVD, MRQSS and SS4.

Under these reverberation conditions, it can be seen that MEDA outperforms the other methods at all SNR levels. This superiority is characterized by a higher NRF coupled with a lower distortion especially at low SNR conditions. Particularly, we note the low performance of MSVD especially at low SNR where it results in a high signal distortion

while offering about 2dB less noise reduction than MEDA. The failure of MSVD in these conditions can be explained by the fact that this method is not capable to cope effectively with this type of interference where the high reverberation time makes the noise field better described as diffuse. As mentioned in Section 6.4.4, MSVD has relatively limited capabilities when the number of noise sources increases [34]. Consequently, in a diffuse noise field, MSVD finds difficulties tracking the noise source, which seems to be impinging from all direction, hence fails to steer a null to the correct noise direction.

MEDA, on the other hand, seems to cope better with these conditions as the assumptions set for its filter design, namely that the noise is uncorrelated on different microphones, are met to an acceptable degree. Indeed, the eigendomain averaging technique appears to considerably boost the performance of MEDA over MRQSS. In fact, in this experiment, as well as in the forthcoming experiments, MEDA systematically outperforms MRQSS both in terms of the noise reduction capabilities and signal distortion, a result which demonstrates the merit of the eigendomain averaging feature which differentiates MEDA from MRQSS.

Experiment 2

In this experiment, the reverberation time is reduced to 100 msec, that is, the noise field is now more accurately described as a coherent noise field with a directional noise source at a DOA of 45 degrees. The results of this experiment are depicted in Figures 8.15 for white noise and 8.16 for colored noise.

It can be observed in these figures that the performance of MSVD has undergone a significant improvement compared to the previous experiment. In fact, the low reverberation allowed this method to detect the direction of the noise more accurately and to successfully eliminate it by steering a null towards it. MEDA, on the other hand, managed to relatively keep the same performance as in the reverberant conditions though not as good as MSVD in this case. It continues nonetheless to outperform MRQSS as expected.

Conducting research to improve MEDA's performance for this kind of interference and make it match that of MSVD is believed to be beneficial. One possible approach could be to exploit the phase in the subvectors of the calculated eigenvectors of the composite covariance matrix (CCM), defined in (6.26). During non-speech activity periods, this phase may be used to estimate the DOA of the noise. Then, instead of the simple conventional beamformer considered here, a more sophisticated design which uses the estimated DOA

to steer a null in the direction of the noise source may be employed. The output of this beamformer is then fed to the MEDA postfilter.

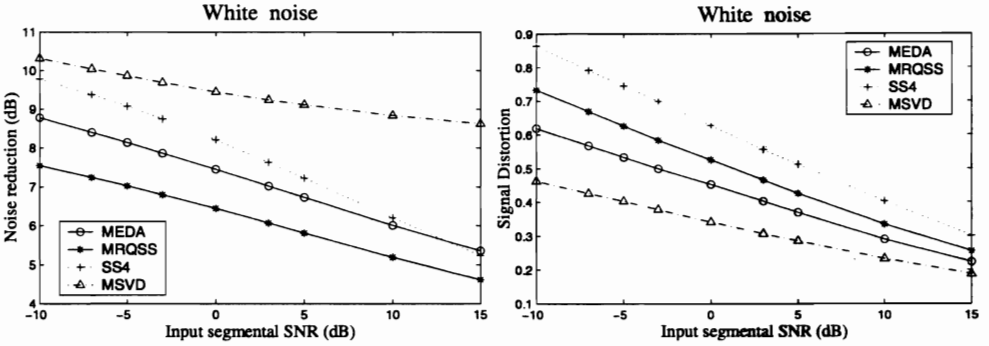


Fig. 8.15 Performance evaluation under different input segmental SNR levels at 100 msec reverberation time for white noise.

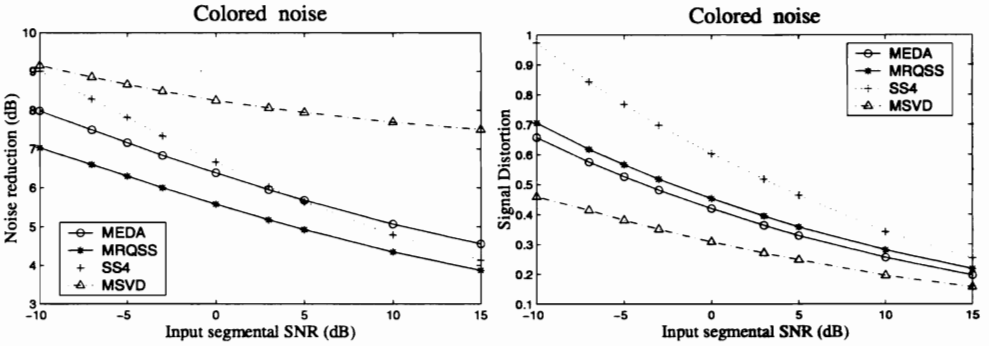


Fig. 8.16 Performance evaluation under different input segmental SNR levels at 100 msec reverberation time for colored noise.

It can also be noted that SS4 shows a better NRF than MEDA but this is achieved at the expense of a non-tolerable signal distortion. In fact this distortion can be reduced by properly adjusting the control parameter μ, which trades off noise reduction to signal distortion, in the gain function. However, this would result in a lower NRF hence the overall performance remains unacceptable. Actually we noticed that the performance of SS4 is largely dependent on μ which is not a desirable feature in practice where it is preferred to maintain an acceptable performance under all or most conditions without any parameter

adjustment. This important property is experimentally found to be met by MEDA[10].

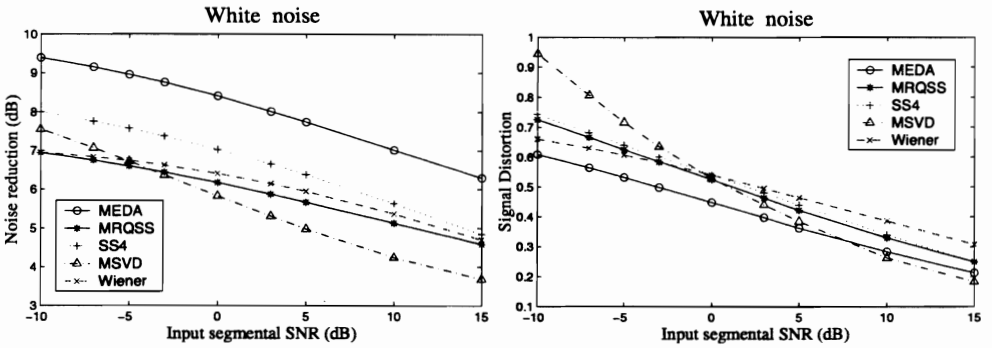


Fig. 8.17 Performance evaluation under different input segmental SNR levels for white noise in the FDRT room.

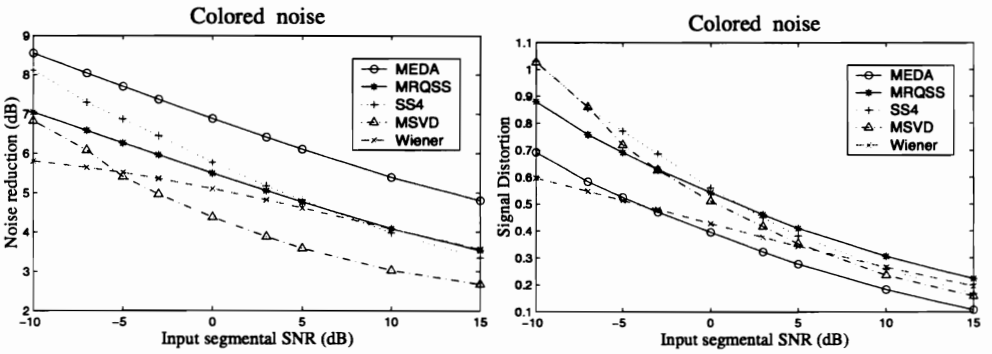


Fig. 8.18 Performance evaluation under different input segmental SNR levels for colored noise in the FDRT room.

Experiment 3

In this third experiment, a more realistic environment is tested. Precisely, the FDRT room is used to simulate an enclosure with frequency dependent reverberation times as shown in Table 8.2.4. Such environment is believed to better simulate realistic room conditions where it is known, as discussed in Chapter 7, that the reflection coefficients, hence the

[10]Actually the chosen value for μ affects the performance but once fixed, the behaviour of MEDA remains relatively stable under most conditions.

reverberation times, are usually dependent on frequency. In this experiment, in addition to the other signal subspace methods, we also evaluate MEDA against the Wiener method. The results of this experiment are shown in Figures 8.17 for white noise and 8.18 for colored noise.

Again the same observations as in the previous experiments can be made in this case. Again MEDA outperforms the other three signal subspace methods and in addition, we also note its superiority over the frequency domain Wiener method. Under these conditions, MEDA exhibits a higher NRF for a similar signal distortion. The Wiener method actually outperforms MRQSS either in terms of signal distortion, noise reduction or both. This again confirms that the superiority of MEDA can mainly be attributed to the eigendomain averaging technique.

8.5.2 Performance versus reverberation time

We next evaluate MEDA as a function of the reverberation time. The room setup described in Section 8.2.4 is again used here. The same reflection coefficient, which also does not depend on frequency, is used for all room surfaces. This coefficient is varied in order to assess the performance of the tested methods as a function of the reverberation time. Two experiments will be performed each with a different input SNR level. In the first experiment, very noisy conditions (SNR = 0 dB) were tested and the results are shown in Figure 8.19 for white noise and Figure 8.20 for colored noise. In the second experiment, 10 dB input SNR conditions were tested. The results are shown in Figure 8.21 and 8.22 for white noise and colored noise respectively.

These experiments confirm the observations made earlier. Indeed it is clearly seen that MEDA relatively maintains the same or slightly better NRF as T_{60} increases. The signal distortion also improves with increasing T_{60}. The MSVD, on the other hand, shows a significant drop in NRF as T_{60} increases. This makes it unsuitable for a diffuse noise field as compared to its superiority under directional interference.

We note also that as far as the signal distortion is concerned, all methods, except SS4, exhibit a moderate increase in distortion until T_{60} is about 200-300 msec then it starts to decrease again. A sound explanation for this result can be as follows. For MSVD, its performance is generally best at low reverberation conditions since the interference can then be regarded as directional noise. As the reverberation time increases, however, MSVD

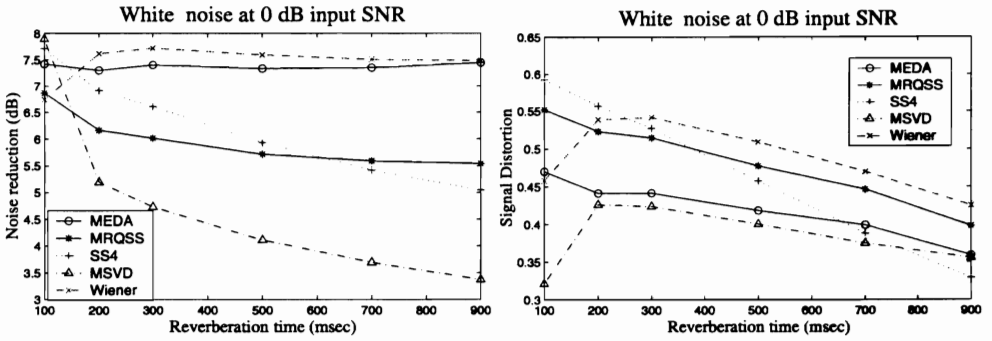


Fig. 8.19 Performance evaluation under different reverberation times for white noise at 0 dB input SNR.

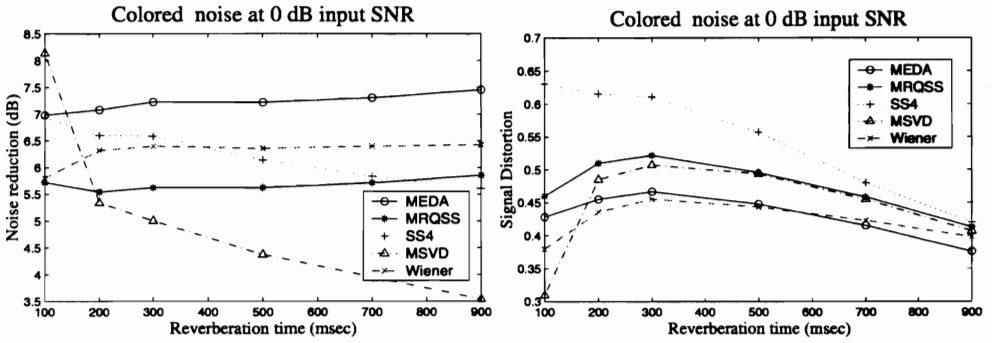


Fig. 8.20 Performance evaluation under different reverberation times for colored noise at 0 dB input SNR.

starts to confuse the noise with the desired speech signal due to reverberation. This results in the observed increase in signal distortion with the suppression filter attempting to cancel more noise. As T_{60} increases further, the noise reduction capabilities of MSVD drastically drops as can be seen in Figures 8.19-8.22. Analogously to the GSC which is known to be no good than a conventional beamformer in diffuse noise fields [8], the MSVD is also expected to behave in a similar manner as it fails to steer the null in the direction of the noise. As less noise cancellation takes place less signal distortion is incurred, which explains the observed shape of the curves.

For the Wiener method, and to a lesser extent MEDA and MRQSS, the opposite scenario occurs. At high reverberation, the noise field is arguably diffuse in which case the

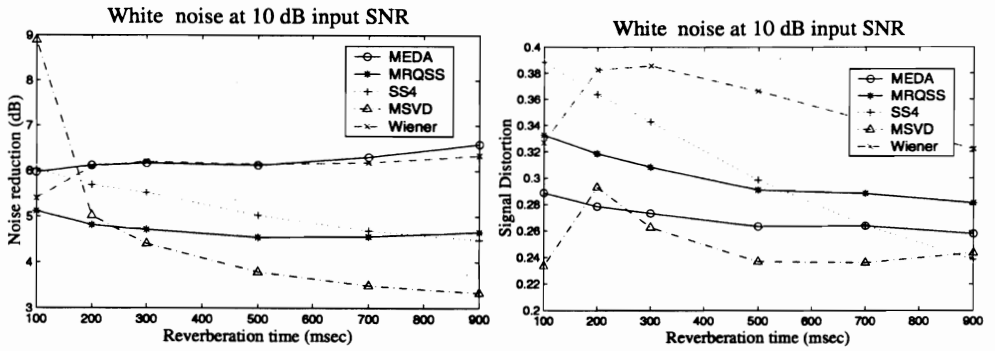


Fig. 8.21 Performance evaluation under different reverberation times for white noise at 10 dB input SNR.

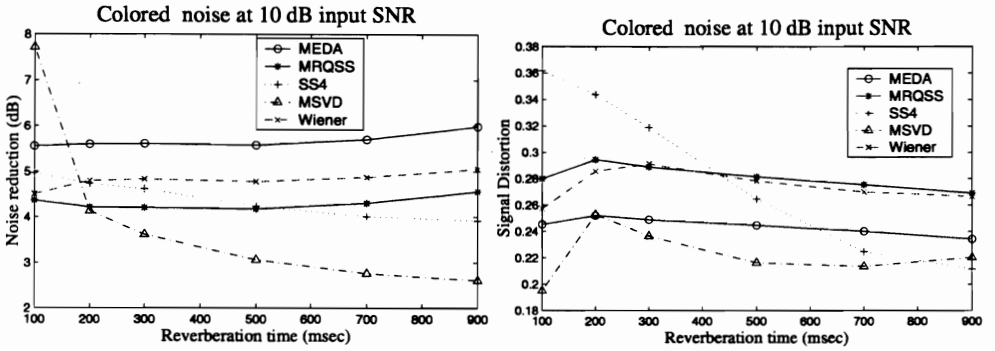


Fig. 8.22 Performance evaluation under different reverberation times for colored noise at 10 dB input SNR.

adaptive postfiltering methods are known to perform best [117]. As the reverberation decreases, the diffuse assumption is gradually violated hence the resulting increase in signal distortion. As the reverberation decreases further the interference becomes more appropriately described as directional in which case adaptive postfiltering methods behave as a conventional beamformer and consequently the distortion decreases with the decrease in noise reduction.

The monotonous decrease in the signal distortion exhibited by SS4 as the reverberation increases is, however, just the result of the less sever noise suppression exerted by the single channel RQSS. Actually taking the average of the individual filter outputs, as given by (8.14), is experimentally found to have little, if any, effect on the resulting signal distortion.

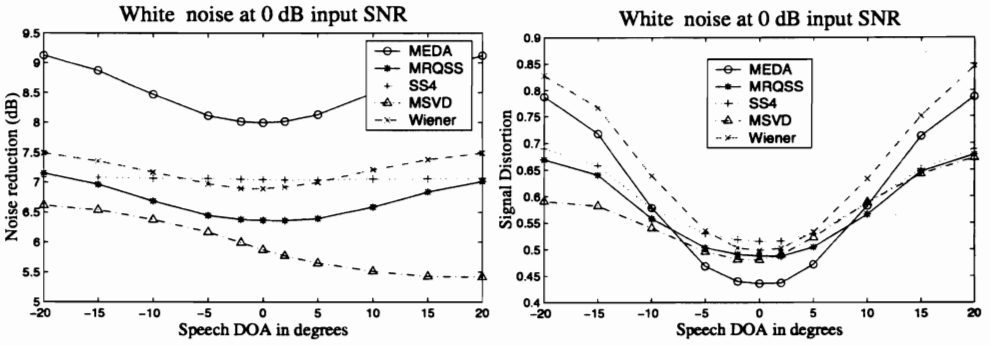


Fig. 8.23 Performance evaluation under different speech DOA in the FDRT room for white noise at 0 dB.

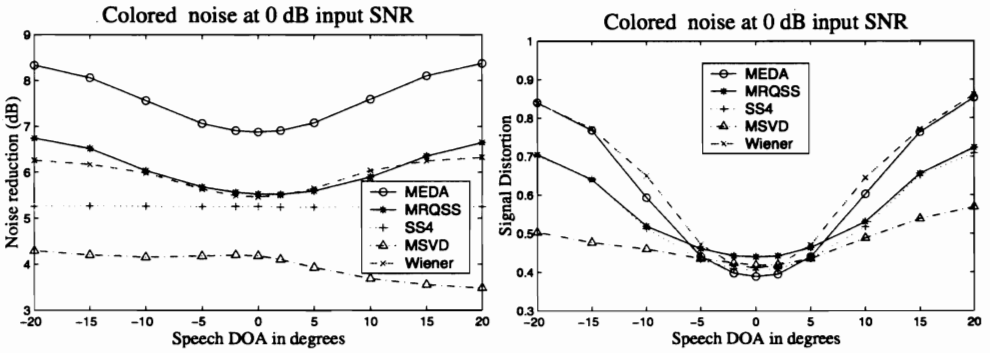


Fig. 8.24 Performance evaluation under different speech DOA in the FDRT room for colored noise at 0 dB.

8.5.3 Sensitivity to steering errors

As discussed in Section 6.6.3, the DOA of the desired speech signal might be a concern in the proposed MEDA method since the filter design is based on the assumption that this signal is perfectly synchronized over all available microphones. For this reason, we test in this experiment the robustness of MEDA against steering errors. To this end, the FDRT room (see Table 8.2.4 for the used reverberation times per frequency band) is considered and the DOA of the desired speech signal is varied while maintaining a fixed distance of 1 m from the center of the microphone array. The noise, on the other hand, is maintained at a fixed DOA of 45° as shown in Figure 8.2. The input segmental SNR is set to 0 dB. The results of these tests are shown in Figures 8.23 and 8.24 for white and colored noises

respectively.

These figures show that the NRF increases when the speech DOA deviates from the assumed 0° broadside direction. However, this increase is actually due to an overall signal cancellation including the suppression of the desired speech signal. Therefore the NRF is not informative in this situation and the signal distortion is the only measure to look at.

Note that for SS4 the NRF is almost constant because the noise suppression is carried out on every channel independently from the other channels. Consequently the noise reduction factor is not affected by the direction of arrival of the desired speech signal. Steering errors, however, do increase the signal distortion.

The MSVD seems to be the least affected by steering errors both in terms of NRF and distortion. These results can be explained by the fact that MSVD makes little assumptions on the direction of the desired speech signal. Note however that as the speech source gets closer to the noise source (positive DOA's) the performance deteriorates.

MEDA on the other hand, is found to be perturbed by the steering errors as expected. However, it still outperforms MRQSS for a misalignment below 5° for colored noise and 10° for white noise. For this reason, we can say that the performance superiority of MEDA can still be maintained if, for example, the steering errors are confined to changes in head position. In fact, for the array and speaker positions shown in Figure 8.2, at a distance of 1 m from the array, a head movement of ± 10 cm will result in a deviation of about $\pm$ 6°. Coupled with a robust time delay compensation module, MEDA is expected to provide satisfactory results. We believe though that further research in this direction should be pursued to improve the overall robustness. A possible solution could be to use the phase shift in the estimated eigenvectors to design a self-calibrating MEDA approach as discussed in Section 6.6.3.

8.6 Conclusion

In this chapter we provided an experimental assessment of the different methods developed in this thesis.

The proposed Frame Based EVD (FBEVD) implementation of the SSA is found to offer considerable computational savings at almost no performance degradation side-effects. This is important since it is believed that the computational issue is the main reason behind the non-popularity of the SSA despite its performance superiority over competing methods.

The FBEVD technique was presented in Section 5.2. Coupling this technique with some form of fast EVD calculation or subspace tracking can give rise to an even more efficient implementation.

After that, we evaluated the novel perceptual signal subspace method (PSS) developed in Chapter 5. It was found, mainly via a series of subjective tests, that PSS provides a perceptually improved performance with a more pleasant speech signal and a less annoying musical noise. Particularly, it was found that the shape of the residual noise remains relatively similar regardless of the original corrupting noise. This result confirms the noise shaping (hence masking) capabilities of PSS. When implemented in conjunction with the FBEVD technique, the added complexity due to the calculation of the masking threshold and its mapping into the eigendomain was insignificant.

Finally, we tested the novel signal subspace multi-microphone method presented in Chapter 6. The MEDA method was found to be a useful speech enhancement tool especially under diffuse noise fields. The performance of MEDA was found to be relatively stable as the reverberation increases. The concern about the sensitivity of MEDA to look direction errors is found to be acceptable and in general it is not more serious than other methods. It was noted that MEDA can still be superior to other methods if the user is instructed to keep a fixed body position while allowed to move his head, even in the absence of a time delay compensation module. Nonetheless, further research can lead to improving MEDA's robustness to steering errors and to directional noise sources.

We note also that during the multi-microphone experiments, the benefit of using the novel subband room simulator, described in Chapter 7, was found to be useful as it significantly reduced the total simulation time.

Chapter 9

Conclusion and Future Work

Speech, by far, remains the most adequate communication tool adopted by humans. For this reason inventions have multiplied over the years offering the customers a more convenient and a more effective use of their preferred communication tool. Supported by the ever growing ambitions and creativity of humans, these systems started to operate under new environments where they are required to offer the same "quality of service" under adverse conditions as they do under quiet. To achieve this goal, speech technology applications emerging on the market nowadays, are in desperate need for sophisticated robust noise cancellation techniques which will allow for a satisfactory operation under the most harsh conditions.

In this thesis, the objective was to build upon the ongoing research efforts and to make a useful contribution to the speech enhancement area. To this end, we investigated the most popular speech enhancement methods and analyzed their advantages and shortcomings. Among the various available techniques and approaches, we were mainly interested in the signal subspace approach (SSA). This choice is mainly motivated by the fact that the SSA is widely considered as a powerful processing tool, for example in the array signal processing area. Besides, the findings reported in the literature of the research conducted so far has revealed that the SSA, when used in speech enhancement applications, outperforms other popular frequency domain techniques.

Actually, the main reason for the non-popularity of the SSA in speech enhancement applications, is the relatively heavy computational load it incurs due to the expensive eigenvalue decomposition (EVD) operation. Fortunately, this handicap is gradually loos-

ing its impact as the recent enormous developments in the DSP technology have offered unprecedented computational power at an affordable low cost.

Nonetheless, we have not ignored this problem in this thesis and the reduction of the computational load of the SSA without significant performance side-effects was one of the issues we addressed. Indeed, we provided a novel implementation scheme for the SSA which we believe may eventually replace the commonly used approach. Our method, called Frame-Based EVD (FBEVD) technique, was experimentally found to considerably reduce the computational burden bringing it to a low cost affordable in real time. The FBEVD simply exploits the stationarity of the speech signal to reduce the rate at which the signal subspace filter is updated. The interesting aspect of this technique is that it accomplishes the desired computational savings without any significant performance degradation.

This technique can open the way to future research in which some fast EVD or subspace tracking techniques can be coupled with the FBEVD. Such a combination, in conjunction with the available high computational power, may lead to efficient robust signal subspace noise reduction methods which can seriously compete with the frequency domain methods, by providing a better tradeoff between performance and computational cost.

The second contribution of this thesis was the incorporation of masking properties of the human ear in the signal subspace approach. The difficulty here arises from the fact that the existing masking models needed to represent the human hearing properties, are generally developed in the the frequency domain. Our work consisted of adopting some signal processing tools making them serve as a mapping between the frequency and the eigen-domains. This mapping, called the Frequency to Eigendomain Transformation (FET), made it possible to represent the perceptual information in the eigendomain leading to a signal subspace filter with improved masking capabilities.

Indeed, a series of subjective tests have revealed the benefit of our Perceptual Signal Subspace (PSS) technique which was designed based on the FET . Particularly, our carefully designed subjective tests showed that PSS achieves a perceptually low residual noise level while maintaining a low signal distortion. One of our tests supported the claim that the residual noise spectral shape remains relatively similar regardless of the original corrupting interference.

One particular issue we encountered during our research was that the masking model we used (as well as other models) are mainly designed for speech coding applications and are not necessarily suitable for speech enhancement. In our specific case, we calculate

the masking threshold based on a low variance, low resolution spectral estimate, namely the Blackman-Tukey estimate. However, since the masking model we used (the MPEG model 1) was intended for a periodogram spectrum, we needed to alter the implementation of the model in order to accommodate it to the encountered differences in the spectral characteristics. While our modifications have indeed resulted in an improved performance, we still believe that putting more effort into the model design would lead to an even better speech enhancement performance.

Moreover, the FET was also used to analyze the SSA from a filterbank standpoint allowing to understand the underlying mechanism from a frequency domain perspective, which is usually more intuitive and instructive for speech signals. The analysis performed confirmed that the eigenvectors of the signal covariance matrix can actually be viewed as filters the pass bands of which usually track the location of the speech formants. The eigenvalues, on the other hand, are the total signal energy at the output of those filters.

The third contribution of this thesis consisted of generalizing the single channel SSA into a microphone array design. The new proposed method is called the Multi-microphone signal subspace method with Eigen-Domain Averaging (MEDA). In this design, we exploit one property of the speech composite covariance matrix (CCM) which carries the spatio-temporal statistics of the signals gathered from the different available microphones. This property states that, under some certain assumptions, an eigenvector of the CCM has equal subvectors. The weighted sum of those subvectors can then be used to estimate the eigenvectors of the speech covariance matrix which span its signal subspace. This eigendomain averaging results in filter coefficients which are more robust to environmental noise leading to an improved speech enhancement performance. The new method is implemented as a conventional beamformer followed by a signal subspace adaptive postfilter.

MEDA has been experimentally found to have a relatively constant performance under different reverberation conditions. This performance can actually slightly improve with increasing reverberation time since that gives rise to a diffuse noise field where the assumptions made by MEDA are better met. Experiments showed that MEDA outperforms the popular frequency domain adaptive postfiltering technique which is known to be suitable for diffuse noise fields. Under a coherent noise field, MEDA manages to maintain a relatively stable performance which is however not as good as methods which can suppress the interference by steering a null towards the noise source.

The experiments in general confirm that the performance superiority of MEDA can

mainly be attributed to the proposed eigendomain averaging technique.

Future research consists in improving the performance of MEDA under directional noise. This can possibly be done by exploiting the observed phase shift in the subvectors of the eigenvectors of the CCM whenever a signal is impinging on the array at an angle different from the one it is steered to. This behaviour can be used in order to estimate the direction of arrival of the interference, during non-speech activity periods, and then design a scheme to cancel the noise impinging from that direction. Besides, these phase shifts can be further investigated in order to conceive a way to build a self-calibrating array in which steering errors are compensated for automatically within the noise reduction algorithm.

Finally, we also developed a novel room simulator which allows to digitally simulate more realistic reverberant enclosures by making it possible to assign frequency dependent wall reflection coefficients in the calculated room impulse responses. The proposed method does so by generalizing the popular image method into a subband implementation. This scheme, by design, readily provides savings in the computational load when computing the simulated reverberated speech signals. This method was particularly useful in this thesis in order to evaluate the performance of the novel MEDA multi-microphone method.

Appendix A

Properties of the matrix C

In this appendix we define and present some of the properties of the matrix $\mathbf{C}$ used in Chapter 6 to simplify the derivations. This matrix is a $MP \times P$ matrix and is defined as follows

$$\mathbf{C} = [\mathbf{I}_P, \ldots, \mathbf{I}_P]^T \tag{A.1}$$

where $\mathbf{I}_P$ is a $P \times P$ identity matrix.

Multiplying a P-dimensional vector $\mathbf{x}$ from the left by $\mathbf{C}$ has the effect of stacking M copies of the vector $\mathbf{x}$ above each other to form an MP-dimensional vector, that is

$$\mathbf{Cx} = \begin{bmatrix} \mathbf{x} \\ \mathbf{x} \\ \vdots \\ \mathbf{x} \end{bmatrix} \tag{A.2}$$

Multiplying a MP-dimensional vector $\mathbf{y} = [\mathbf{y}_1^T, \ldots, \mathbf{y}_M^T]^T$ from the left by $\mathbf{C}^T$, where $\mathbf{y}_i$'s are P dimensional sub-vectors, adds up these sub-vectors in the following way,

$$\mathbf{C}^T\mathbf{y} = \sum_{i=1}^{M} \mathbf{y}_i \tag{A.3}$$

Therefore using (A.2) and (A.3) it can easily be seen that

$$\text{If } \mathbf{y} = \mathbf{Cx} \text{ then } \mathbf{x} = \frac{1}{M}\mathbf{C}^T\mathbf{y} \tag{A.4}$$

The reciprocal, however, is not necessarily true.

These results can be extended to matrices in the following way. Consider the $P \times P$ matrix $\mathbf{B}$ then we have,

$$\mathbf{C}\mathbf{B}\mathbf{C}^T = \begin{bmatrix} \mathbf{B} & \cdots & \mathbf{B} \\ \vdots & \ddots & \vdots \\ \mathbf{B} & \cdots & \mathbf{B} \end{bmatrix} \tag{A.5}$$

that is $\mathbf{C}\mathbf{B}\mathbf{C}^T$ is a $MP \times MP$ matrix equal to $\mathbb{I}_M \otimes \mathbf{B}$ where $\mathbb{I}_M$ is an $M \times M$ all ones matrix and $\otimes$ is the Kronecker product [50].

In a similar way, and defining the $MP \times MP$ matrix $\mathbf{A}$ as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \cdots & \mathbf{A}_{1M} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{M1} & \cdots & \mathbf{A}_{MM} \end{bmatrix}$$

then we have

$$\mathbf{C}^T\mathbf{A}\mathbf{C} = \sum_{i=1}^{M}\sum_{j=1}^{M} \mathbf{A}_{ij} \tag{A.6}$$

Finally using (A.6) and (A.5) respectively we obtain

$$\mathbf{C}^T\mathbf{C} = \mathbf{C}^T\mathbf{I}_{MP}\mathbf{C} = M\mathbf{I}_P \tag{A.7}$$

and

$$\mathbf{C}\mathbf{C}^T = \mathbf{C}\mathbf{I}_P\mathbf{C}^T = \mathbb{I}_M \otimes \mathbf{I}_P \tag{A.8}$$

Appendix B

Detailed MEDA Experimental Results

In Section 8.5, all the experimental results of the multi-microphone methods were given for white and colored noises. While the white noise was computer generated, the colored noise results were obtained as the average performance of four colored noise types. Namely, a kitchen fan a computer fan, a dryer and a freezer motor noise. For the interested reader, he detailed results of these noises are given in this appendix for all the experiments described.

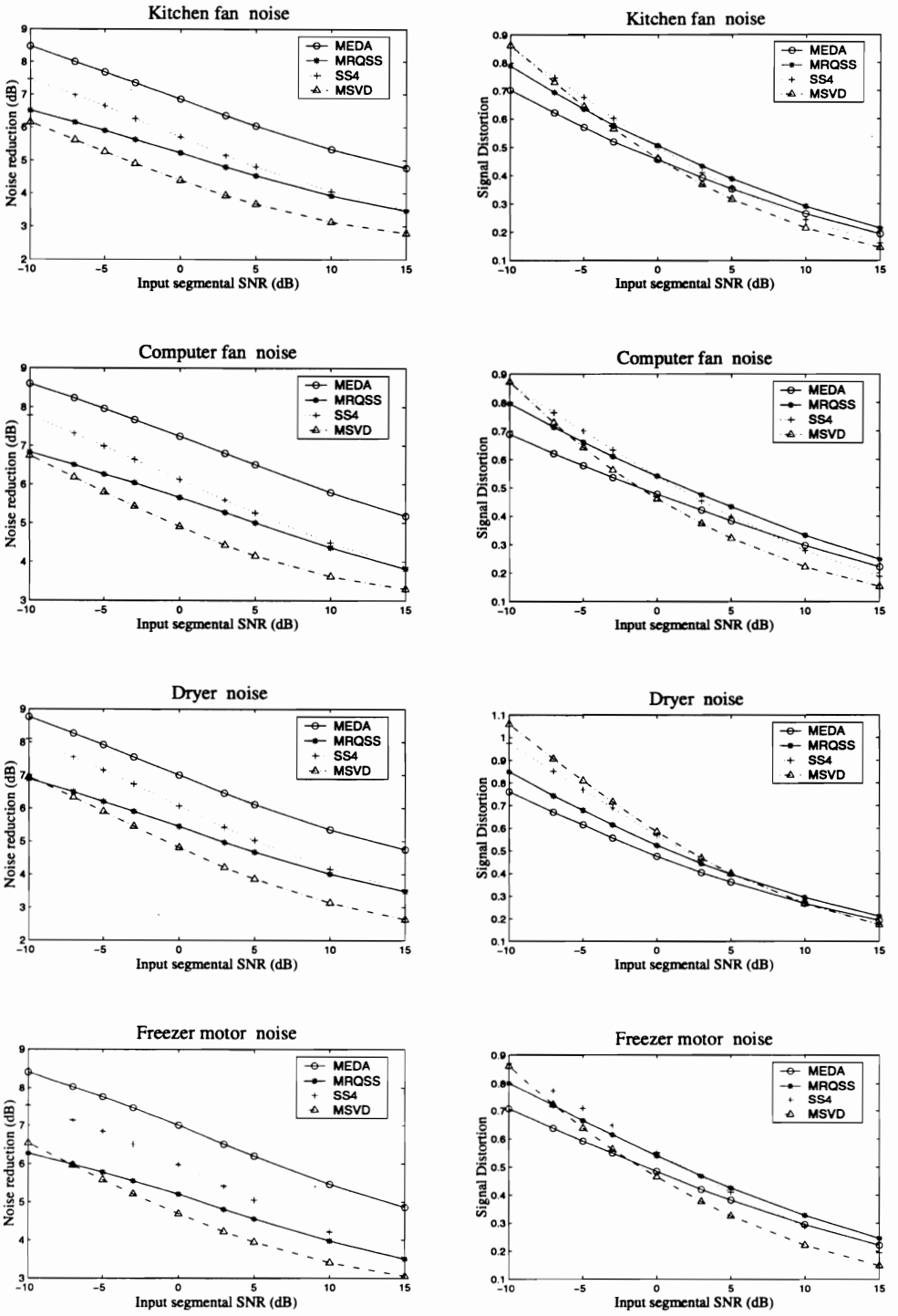


Fig. B.1 Performance evaluation under different input segmental SNR levels at 400 msec reverberation time for various colored noises.

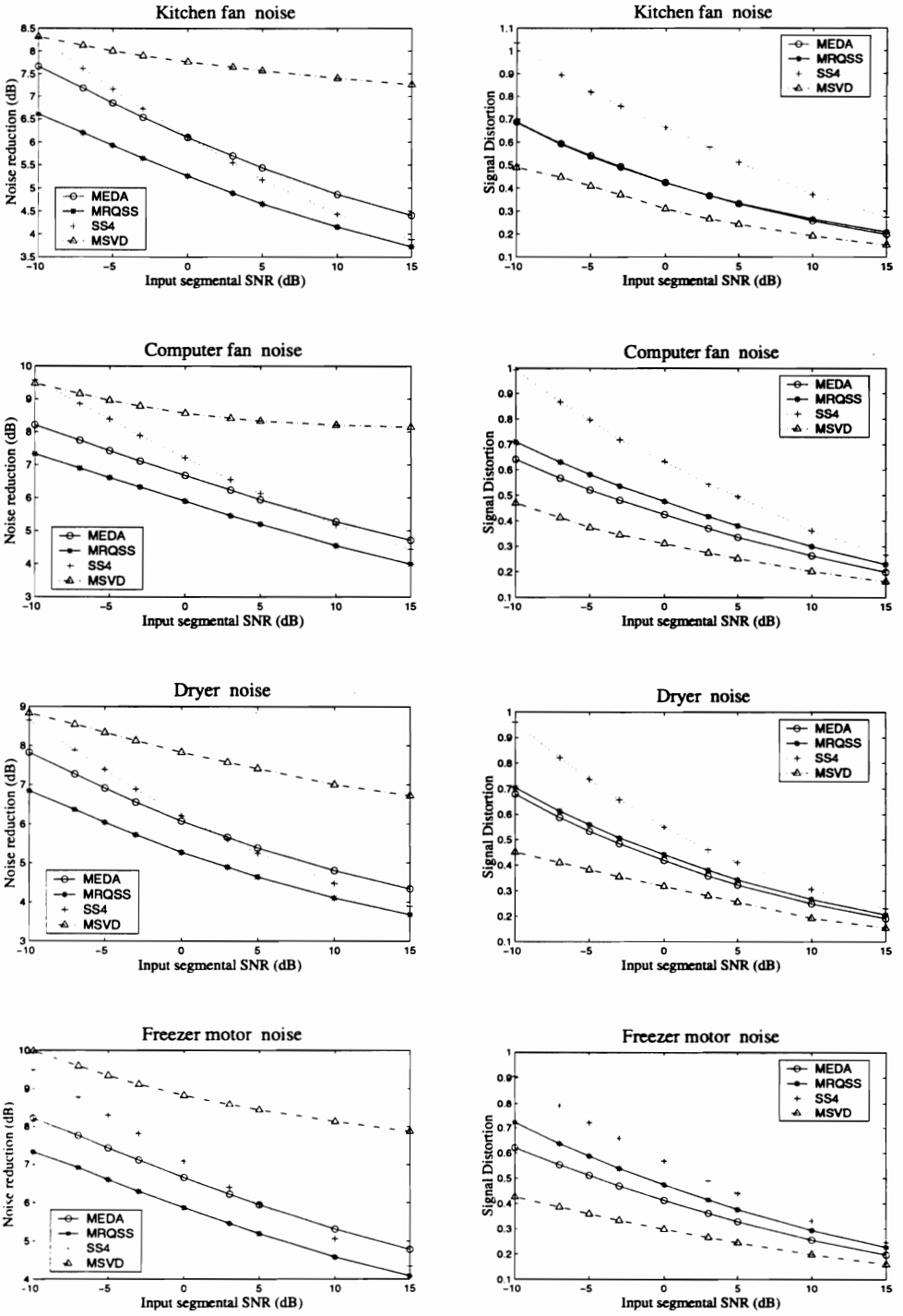


Fig. B.2 Performance evaluation under different input segmental SNR levels at 100 msec reverberation time for various colored noises.

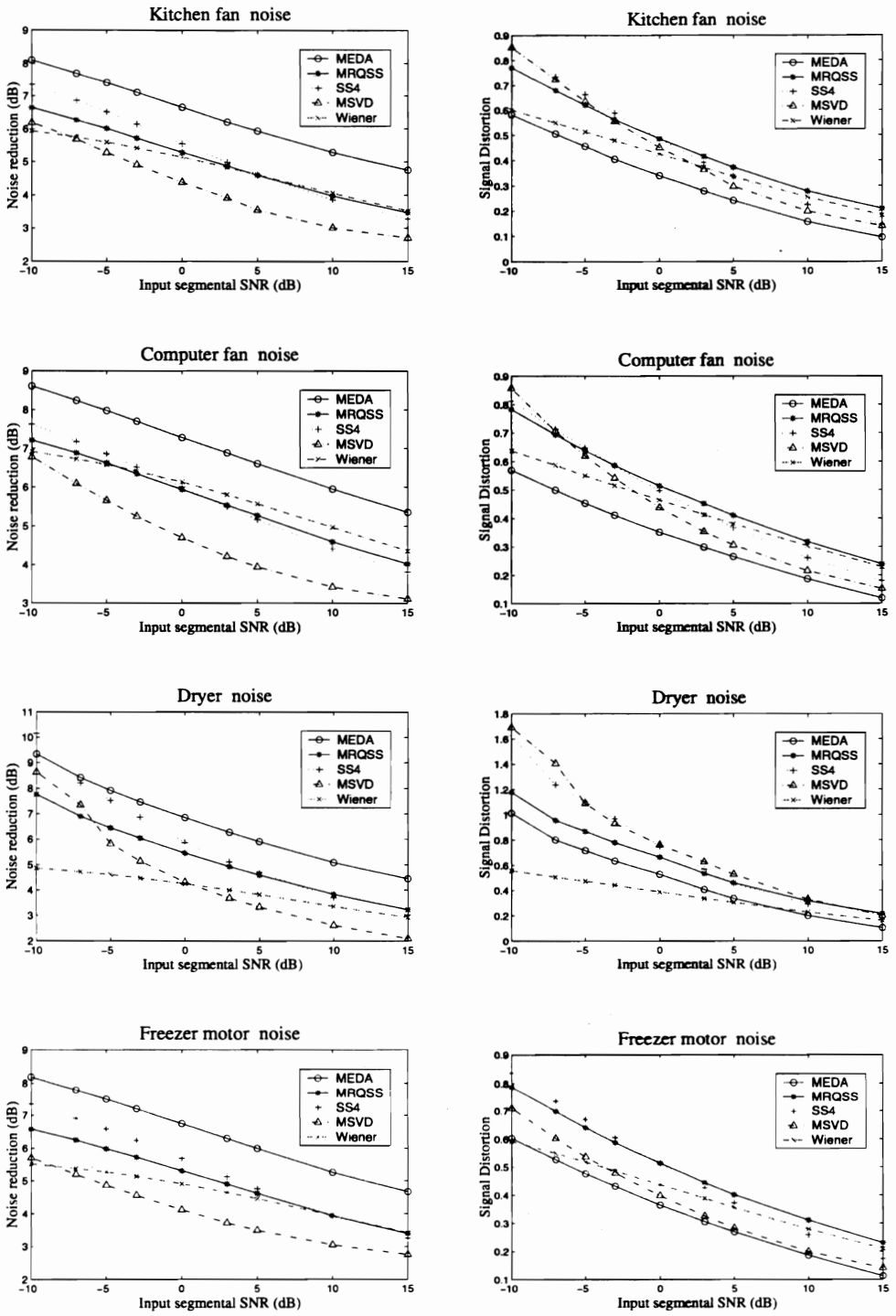


Fig. B.3 Performance evaluation under different input segmental SNR levels in the FDRT room for various colored noises.

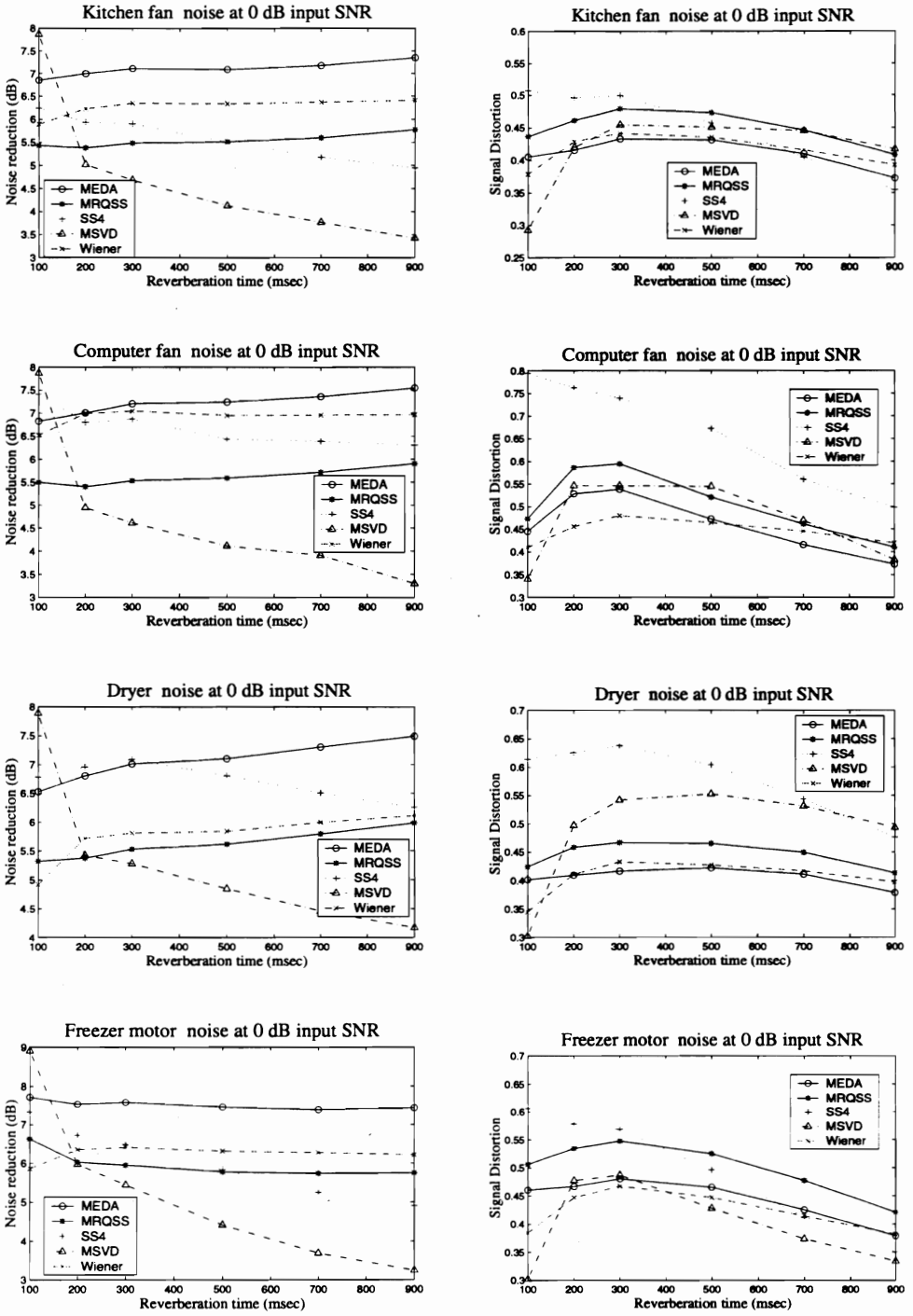


Fig. B.4 Performance evaluation as a function of reverberation times for different types of colored noise at 0 dB input SNR.

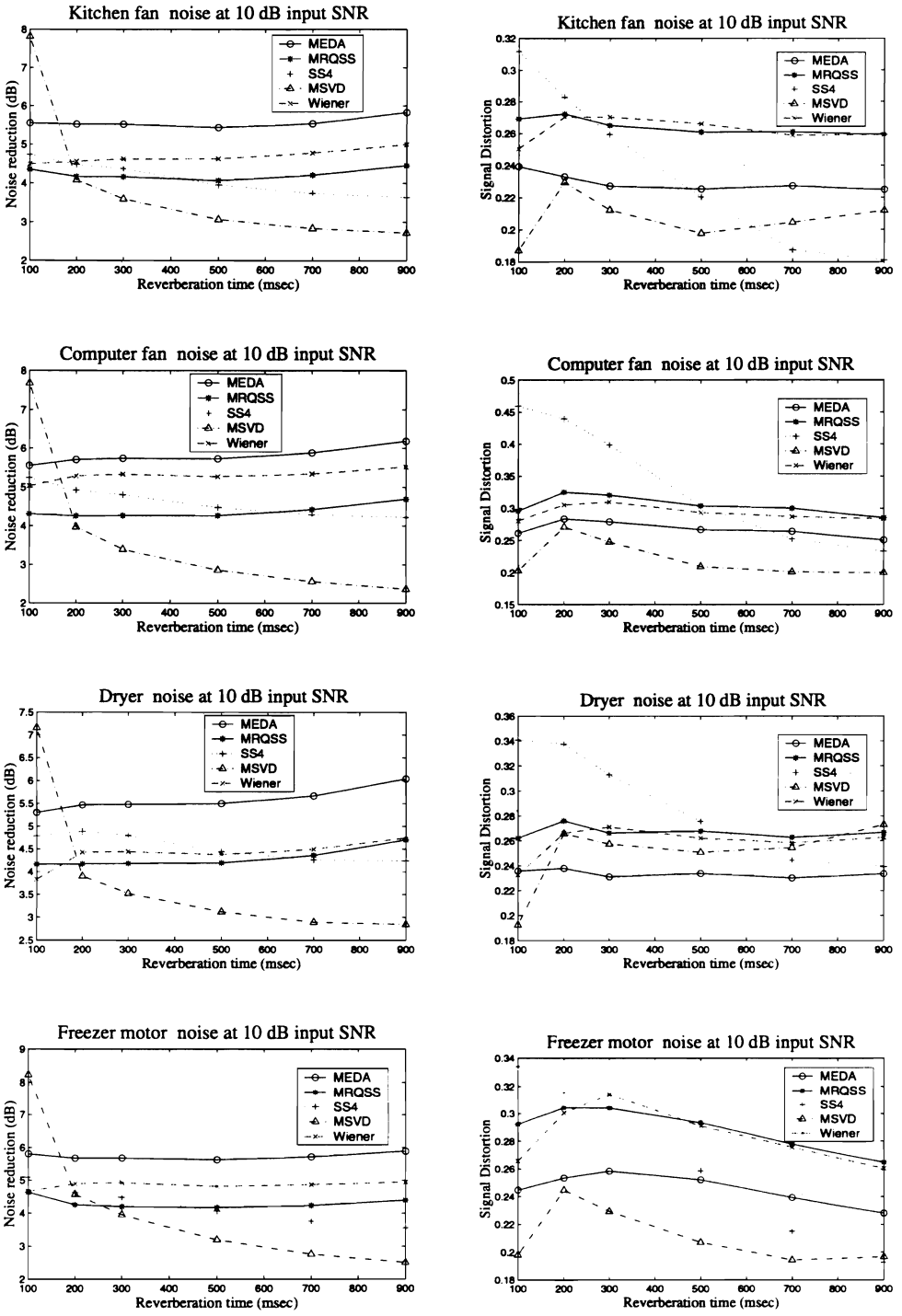


Fig. B.5 Performance evaluation as a function of reverberation times for different types of colored noise at 10 dB input SNR.

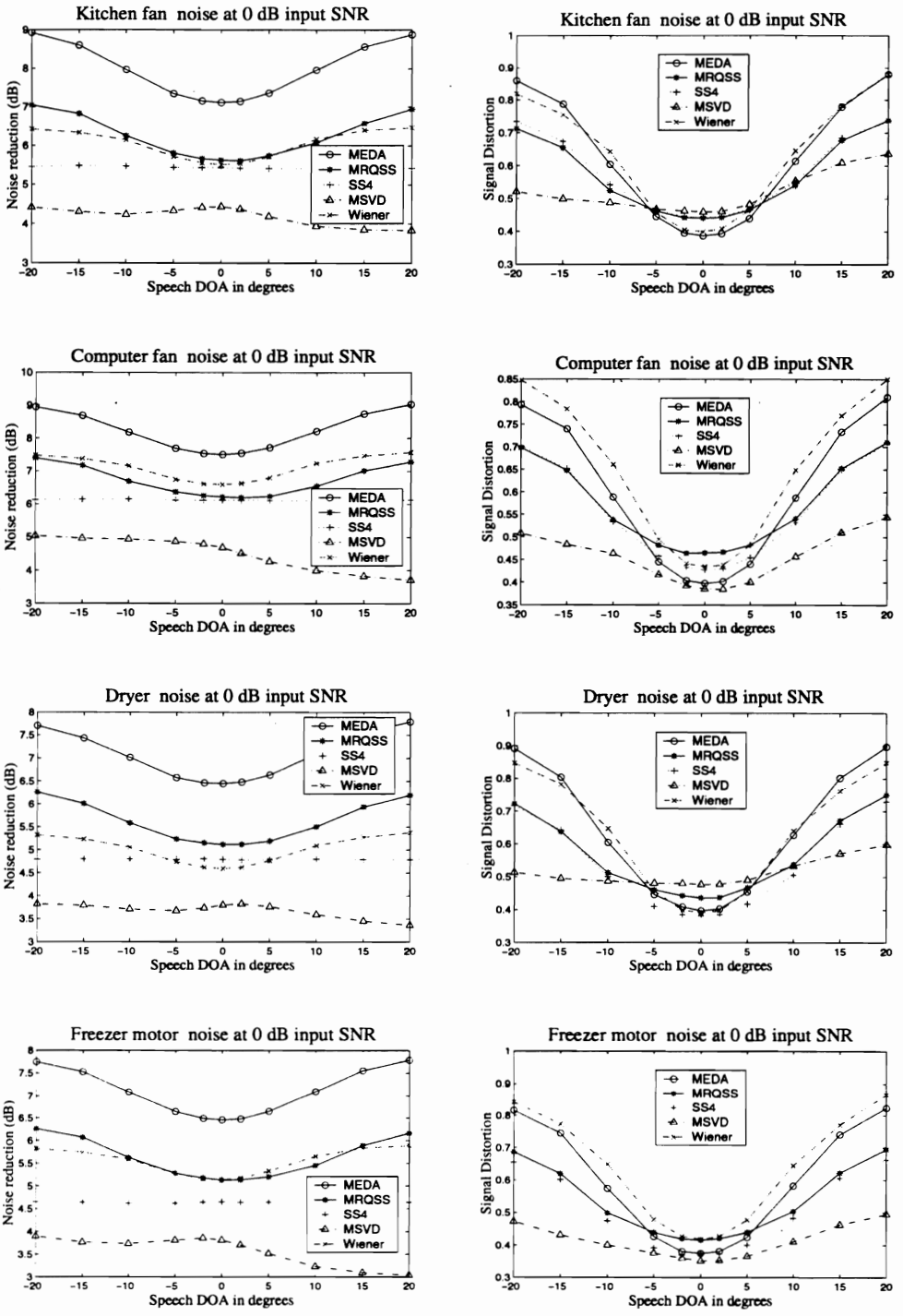


Fig. B.6 Performance evaluation under different speech DOA in the FDRT room for different colored noises at 0 dB.

References

[1] S. Affes and Y. Grenier. A signal subspace tracking algorithm for microphone array processing of speech. *IEEE Trans. on Speech and Audio Processing*, 5(5):425–437, September 1997.

[2] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Amer.*, 65(4):943–950, April 1979.

[3] J. B. Allen, D. A Berkley, and J. Blauert. Multimicrophone signal-processing technique to remove room reverberation from speech signals. *J. Acoust. Soc. Amer.*, 62(4):912–915, 1997.

[4] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura. Speech enhancement based on the subspace method. *IEEE Trans. on Speech and Audio Processing*, 8(5):497–507, September 2000.

[5] A. A. Azirani, R. J. Le Bouquin, and G. Faucon. Optimizing speech enhancement by exploiting masking properties of the human ear. *in Proc ICASSP95*, pages 800–803, 1995.

[6] J. G. Beerends and J. A. Stemerdink. A perceptual audio quality measure based on psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 40:963–978, December 1992.

[7] M. Berouti, R. Schwarz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. *in Proc ICASSP79*, 1:208–211, 1979.

[8] J. Bitzer, K. U. Simmer, and K. D. Kammeyer. Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement. *in Proc ICASSP99*, 5, 1999.

[9] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 27(2):113–120, April 1979.

[10] S. F. Boll and R. E. Wohlford. Event driven speech enhancement. *in Proc ICASSP83*, pages 1152–1155, 1983.

[11] S. E. Bou-Ghazale and K. Assaleh. A robust endpoint detection of speech for noisy environments with application to automatic speech recognition. *in Proc ICASSP02*, 4:3808–3811, 2002.

[12] K. Brandenburg and G. Stoll. ISO-MPEG-1 Audio: A generic standard for coding of high quality digital audio. *Journal of the Audio Engineering Society*, 42(10):780–792, October 1994.

[13] M. S. Brandstein, J. E. Adcock, and H. F. Silverman. A localization-error-based method for microphone-array design. *in Proc ICASSP97*, 1:375–378, 1997.

[14] M. S. Brandstein and H. F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. *in Proc ICASSP96*, 2:901–904, 1996.

[15] Y. Bresler and A. Macovski. Exact maximum likelihood parameter estimation of superimposed exponential signals in noise. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 34(5):1081–1089, October 1986.

[16] G. C. Carter. *Coeherence and Time Delay Estimation.* IEEE Press, Piscataway, NJ, 1993.

[17] B. Champagne. Simulation of the response of multiple microphones to a moving point source. *Applied Acoustics*, 42:313–332, 1994.

[18] B. Champagne, S. Bedard, and A. Stephenne. Performance of time delay estimation in the presence of room reverberation. *IEEE Trans. on Speech and Audio Processing*, 4:148–152, February 1996.

[19] B. Champagne and Q. G. Liu. Plane rotation-based EVD updating schemes for efficient subspace tracking. *IEEE Trans. on Signal Processing*, 46(7):1886–1900, July 1998.

[20] T. Chou. Frequency-independent beamformer with low response error. *in Proc ICASSP95*, pages 2995–2998, 1995.

[21] P. L. Chu. Desktop mic array for teleconferencing. *in Proc ICASSP95*, 5:2999–3002, 1995.

[22] I. Cohen and B. Berdugo. Noise estimation by minima controlled recursive averaging for speech enhancement. *IEEE Signal Processing letters*, 9(1):12–15, January 2002.

[23] C. Colomes, M. Lever, J. B. Rault, Y. F. Dehery, and G. Faucon. A perceptual model applied to audio bit rate reduction. *Journal of the Audio Engineering Society*, 43(4):233–240, April 1995.

[24] R. E. Crochiere and L. R. Rabiner. *Multirate Digital Signal Processing.* Prentice-Hall, Englewood Cliffs, NJ, 1983.

[25] A. Czyzewski and R. Krolikowski. Noise reduction in audio signals based on the perceptual coding approach. *Proc IEEE WASPAA*, pages 147–150, 1999.

[26] M. Dahl, I. Claesson, and S. Nordebo. Simultaneous echo cancellation and car noise suppression employing a microphone array. *IEEE Trans. on Vehicular Technology*, 48(5):1518–1526, September 1999.

[27] T. Dau, D. Puschel, and A. Kohlrausch. A quantative model of the effective signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Amer.*, 99(6):3615–3622, June 1996.

[28] T. Dau, D. Puschel, and A. Kohlrausch. A quantative model of the effective signal processing in the auditory system. II. simulations and measurements. *J. Acoust. Soc. Amer.*, 99(6):3623–3631, June 1996.

[29] J. M. De Haan, N. Grbic, I. Claesson, and S. Nordholm. Design of oversampled uniform dft filter banks with delay specification using quadratic optimization. *in Proc ICASSP01*, 6:3633 –3636, 2001.

[30] N. D. Degan and C. Parti. Acoustic noise analysis and speech enhancement techniques for mobile radio applications. *Signal Processing*, 15:43–56, July 1988.

[31] J. R. Deller, Proakis J. G, and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals.* Prentice-Hall, Englewood Cliffs, NJ, 1993.

[32] M. Dendrinos, S. Bakamidis, and G. Carayannis. Speech enhancement from noise: A regenerative approach. *Speech Communications*, 10(1):45–57, February 1991.

[33] G. Doblinger. Computationally efficient speech enhancement by spectral minima tracking in subbands. *Proc. Eurospeech*, 2:1513–1516, 1995.

[34] S. Doclo. *Multi-Microphone Noise Reduction and Dereverberation Techinques for Speech Applications.* PhD thesis, Katholieke Universiteit Leuven, Heverlee, Belgium, May 2003.

[35] S. Doclo and M. Moonen. GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Trans. on Signal Processing*, 50(9):2230–2244, September 2002.

[36] S. Doclo, M. Moonen, and E. De Clippel. Combined acoustic echo and noise reduction using GSVD-based optimal filtering. *in Proc ICASSP00*, 2:1061–1064, 2000.

[37] A. Dutta-Roy. Virtual meetings with desktop conferencing. *IEEE Spectrum*, 35(7):47–56, July 1998.

[38] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 32:1109–1121, December 1984.

[39] Y. Ephraim, D. Malah, and B. H. Juang. On the applications of Hidden Markov Models for enhancing noisy speech. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 37:1864–1865, December 1989.

[40] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *in Proc ICASSP93*, 2:355–358, 1993.

[41] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. on Speech and Audio Processing*, 3(4):251–266, July 1995.

[42] N. W. D. Evans and J. S. Mason. Noise estimation without explicit speech, non-speech detection: a comparison of mean, median and modal based approaches. *Proc. Eurospeech*, 2:893–896, 2001.

[43] N. W. D. Evans, J. S. Mason, and B. Fauve. Efficient realtime noise estimation without explicit speech, non-speech detection: An assessment on the aurora corpus. *Proc. IEEE Int. Conf. DSP*, page 985988, 2002.

[44] J. Flanagan, D. Berkley, G. Elko, J. West, and M. Sondhi. Autodirective microphone systems. *Acustica*, 73:58–71, 1991.

[45] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Eiko. Computer-steered microphone arrays for sound transduction in large rooms. *J. Acoust. Soc. Amer.*, 78(5):1508–1518, November 1985.

[46] H. Fletcher. Auditory patterns. *Rev. Mod. Phys.*, 12:47–65, 1940.

[47] D. A. Florencio and H. S. Malvar. Multichannel filtering for optimum noise reduction in microphone arrays. *in Proc ICASSP01*, 1:197–200, 2001.

[48] S. Ghahramani. *Fundamentals of Probability.* Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 2000.

[49] A. Gilloire and M. Vetterli. Adaptive filtering in subbands with critical sampling: analysis, experiments and application to acoustic echo cancellation. *IEEE Trans. on Signal Processing*, 40:1862–1875, August 1992.

[50] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, MD, 2nd edition, 1989.

[51] M. M. Goodwin and G. W. Elko. Constant beamwidth beamforming. *in Proc ICASSP93*, 1:169–172, 1993.

[52] M. M. Goulding and J. S. Bird. Speech enhancement for mobile telephony. *IEEE Trans. on Vehicular Technology*, 39(4):316–326, November 1990.

[53] R. M. Gray. On the asymptotic eigenvalue distribution of toeplitz matrices. *IEEE Trans. on Info. Theory*, 18(6):725–730, November 1972.

[54] Y. Grenier. A microphone array for car environments. *Speech Communications*, 12(1):25–39, March 1993.

[55] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. on Antennas and Propagation*, 30(1):27–34, January 1982.

[56] S. Gustafsson, P. Jax, and P. Vary. A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics. *in Proc ICASSP98*, pages 397–400, 1998.

[57] S. Gustafsson, R. Martin, P. Jax, and P. Vary. A psychoacoustic approach to combined acoustic echo cancellation and noise reduction. *IEEE Trans. on Speech and Audio Processing*, 10(5):245–256, July 2002.

[58] T. Gustafsson, B. D. Rao, and M. Trivedi. Analysis of time-delay estimation in reverberant environments. *in Proc ICASSP02*, 2:2097–2100, 2002.

[59] J. H. L. Hansen and M. A. Clements. Constrained iterative speech enhancement with application to speech recognition. *IEEE Trans. on Signal Processing*, 39(4):795–805, April 1991.

[60] P. C. Hansen and S. H. Jensen. Fir filter representation of reduced-rank noise reduction. *IEEE Trans. on Signal Processing*, 46(6):1737–1741, June 1998.

[61] P. S. K. Hansen. *Signal Subspace Methods for Speech Enhancement.* PhD thesis, Technical University of Denmark, Lyngby, Denmark, September 1997.

[62] P. S. K. Hansen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen. Noise reduction of speech signals using the rank-revealing ULLV decomposition. *EUSIPCO96*, pages 182–185, 1996.

[63] M. Harteneck, S. Weiss, and R. W. Stewart. Design of near perfect reconstruction oversampled filter banks for subband adaptive filters. *IEEE Trans. on Circuits and Systems II: Analog and Digital Signal Processing*, 46(8):1081 –1085, August 1999.

[64] M. H. Hayes. *Statistical Digital Signal Processing and Modeling.* John Wiley & Sons, Inc, New York, 1996.

[65] S. Haykin. *Adaptive Filter Theory.* Prentice-Hall, Englewood Cliffs, NJ, 4th edition, 2002.

[66] H. G. Hirsch and C. Ehrlicher. Noise estimation techniques for robust speech recognition. *in Proc ICASSP95*, 1:153–156, 1995.

[67] M. P. Hollier, M. O. J. Hawksford, and D. R. Guard. Algorithms for assessing the subjectivity of perceptually weighted audible errors. *Journal of the Audio Engineering Society*, 43(12):1041–1045, December 1995.

[68] Y. Hu and C. Loizou. A subspace approach for enhancing speech corrupted by colored noise. *in Proc ICASSP02*, 1:573–576, 2002.

[69] J. Huang and Y. Zhao. An energy-constrained signal subspace method for speech enhancement and recognition in colored noise. *Speech Communications*, 1:165–181, 1998.

[70] J. Huang and Y. Zhao. An energy-constrained signal subspace method for speech enhancement and recognition in white and colored noise. *in Proc ICASSP98*, 1:377–380, 1998.

[71] J. Huang and Y. Zhao. A DCT-based fast signal subspace technique for robust speech recognition. *IEEE Trans. on Speech and Audio Processing*, 8(6):747–751, November 2000.

[72] International Satandards Organization. ISO/IEC JTC1/SC29/WG 11. *Coding of Moving Pictures and Associated Audio*, April 1993.

[73] International Satandards Organization. ISO/IEC DIS 13818-7. *Generic Coding of Moving Pictures and Associated Audio Information (Part 7)-Advanced Audio Coding (AAC)*, 1996.

[74] ITU-R BS.1387 Recommendation. Method for objective measurements of perceived audio quality. *International Telecommunication Union*, 1998.

[75] ITU-T P.862 Recommendation. Perceptual evaluation of speech quality (PESQ). *International Telecommunication Union*, February 2001.

[76] F. Jabloun. A signal subspace approach for speech enhancement using masking properties of the human ear. *in ICASSP01, student Forum, abstract only*, 2001.

[77] F. Jabloun and B. Champagne. A fast subband room response simulator. *in Proc ICASSP00*, 2:925–928, 2000.

[78] F. Jabloun and B. Champagne. A multi-microphone signal subspace approach for speech enhancement. *in Proc ICASSP01*, 1:205–208, 2001.

[79] F. Jabloun and B. Champagne. On the use of masking properties of the human ear in the signal subspace speech enhancement approach. *in Proc IWAENC, Darmstadt*, pages 199–202, 2001.

[80] F. Jabloun and B. Champagne. A perceptual signal subspace approach for speech enhancement in colored noise. *in Proc ICASSP02*, 1:569–572, 2002.

[81] F. Jabloun and B. Champagne. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 11(6):700–708, November 2003.

[82] N. Jayant, J. Johnston, and R. Safranek. Signal compression based on models of human perception. *Proc. IEEE*, 81(10):1385–1422, October 1993.

[83] J. Jensen and J. H. L. Hansen. Speech enhancement using a constrained iterative sinusoidal model. *IEEE Trans. on Speech and Audio Processing*, 9(7):731–740, October 2001.

[84] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen. Reduction of broadband noise in speech by truncated QSVD. *IEEE Trans. on Speech and Audio Processing*, 3(6):439–448, November 1995.

[85] M. Jeppesen, C. A. Rodbro, and S. H. Jensen. Recursively updated eigenfilterbank for speech enhancement. *in Proc ICASSP01*, 1:653–656, 2001.

[86] D. Johnson and D. Dudgeon. *Array signal processing, Concepts and Techniques.* New Jersey: Prentice Hall, 1st edition, 1993.

[87] J. Johnston. Second generation audio coding: The hybrid coder. *88th AES Convention, Preprint 2937*, March 1990.

[88] J. D. Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE J Select. Areas Commun*, 6(2):314–323, February 1988.

[89] J. C. Junqua, B. Mak, and B. Reaves. A robust algorithm for word boundary detection in the presence of noise. *IEEE Trans. on Speech and Audio Processing*, 2(3):406–412, July 1994.

[90] J. C. Junqua, B. Reaves, and B. Mak. A study of of endpoint detection algorithms in adverse conditions: Incidence on a dtw and hmm recognizer. *Proc. Eurospeech 91*, pages 1371–1374, 1991.

[91] Y. Kaneda. Adaptive microphone array system for noise reduction (AMNOR) and its performance studies. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 34:1391–1400, December 1986.

[92] Y. Kaneda and J. Ohga. Adaptive microphone array system for noise reduction. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 34(6):1391–1400, 1986.

[93] J. Kates. A comparaison of hearing aid array processing techniques. *J. Acoust. Soc. Amer.*, 99(5):3138–3148, May 1996.

[94] S. M. Kay. *Modern Spectral Estimation Theory and Applications.* Prentice Hall, 1988.

[95] W. Kellermann. A self-steering digital microphone array. *in Proc ICASSP91*, 5:3581 –3584, 1991.

[96] F. Khalil, J. Jullien, and A. Gilloire. Microphone array for sound pickup in teleconference systems. *Journal of the Audio Engineering Society*, 42(9):691–700, 1994.

[97] J. U. Kim, S. G. Kim, and C. D. Yoo. The incorporation of masking threshold to subspace speech enhancement. *in Proc ICASSP03*, 2003.

[98] K. Kiyohara, Y. Kaneda, S. Takahashi, H. Nomura, and J. Kojima. A microphone array system for speech recognition. *in Proc ICASSP97*, 1:215–218, 1997.

[99] M. Klein and P. Kabal. Signal subspace speech enhancement with perceptual post-filtering. *in Proc ICASSP02*, 1:537–540, 2002.

[100] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 24(4):320–326, August 1976.

[101] H. Krim and M. Viberg. Two decades of array signal proceesing research: The parametric approach. *IEEE Signal Processing Magazine*, pages 67–94, July 1996.

[102] J. Kumagai. Talk to the machine. *IEEE Spectrum*, pages 6–064, September 2002.

[103] H. Kuttruff. *Room Acoustics.* 3rd Edn. Elsevier, London, 1991.

[104] R. Le Bouquin and G. Faucon. Using the coherence function for noise reduction. *IEE Proceedings-I Commun., Speech, Vision*, 139:276–280, June 1992.

[105] R. Le Bouquin-Jeannes, A. A. Azirani, and G. Faucon. Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator. *IEEE Trans. on Speech and Audio Processing*, 5(5):484–487, November 1997.

[106] T. S. Lee. Efficient wideband source localization using beamforming invariance technique. *IEEE Trans. on Signal Processing*, 42(6):1376–1387, June 1994.

[107] J. S. Lim. Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 26(5):471–472, October 1978.

[108] J. S. Lim and A. V. Oppenheim. All-Pole modeling of degraded speech. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 26:197–210, June 1978.

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

[110] Q. Liu, B. Champagne, and D. K. C. Ho. Simple design of oversampled uniform DFT filter banks with applications to subband acoustic echo cancellation. *Signal Processing*, 80(5):831–847, May 2000.

[111] Q. G. Liu and B. Champagne. A microphone array processing technique for speech enhancement in a reverberant space. *Speech Communications*, 18:317–334, 1996.

[112] P. Lockwood and J. Boudy. Experiments with a nonlinear spectral subtractor (NSS), hidden markov models and the projection, for robust speech recognition in cars. *Speech Communication*, 11:215–228, 1992.

[113] D. G. Luenberger. *Linear and Nonlinear Programming.* Addison-Wesley, Reading, MA, 1984.

[114] D. Mahmoudi and A. Drygajlo. Combined Wiener and coherence filtering in wavelet domain for microphone array speech enhancement. *in Proc ICASSP98*, 1:385–388, 1998.

[115] R. J. Mailloux. *Phased Array Antenna Handbook.* Artech House, Boston, 1993.

[116] S. L. Marple. *Digital Spectral Analysis with Applications.* Prentice Hall, 1987.

[117] C. Marro, Y. Mahieux, and K. U. Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Trans. on Speech and Audio Processing*, 6(3):240–259, May 1998.

[118] R. Martin. Spectral subtraction based on minimum statistics. *Proc. EUSIPCO*, pages 1182–1185, 1994.

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

[120] J. Meyer and K. U. Simmer. Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction. *in Proc ICASSP97*, 2:1167–1170, 1997.

[121] U. Mittal and N. Phamdo. Signal/Noise KLT based approach for enhancing speech degraded by colored noise. *IEEE Trans. on Speech and Audio Processing*, 8(2):159–167, March 2000.

[122] M. Omologo and P. Svaizer. Use of the crosspower-spectrum phase in acoustic event location. *IEEE Trans. on Speech and Audio Processing*, 5(3):288–292, May 1997.

[123] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-time Signal Processing.* Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1999.

[124] D. O'Shaughnessy. *Speech Communications Human and Machine.* IEEE Press, New York, 2nd edition, 2000.

[125] B. Paillard, P. Mabilleau, S. Morissette, and J. R. Soumagne. PERCEVAL: Perceptual evaluation of the quality of audio signals. *Journal of the Audio Engineering Society*, 40:21–31, Jan./Feb 1992.

[126] P. M. Peterson. Simulating the response of multiple microphones to a single acoustic source in a reverberant room. *J. Acoust. Soc. Amer.*, 80(5):1527–1529, November 1986.

[127] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements. *Objective measures of Speech Quality.* Prentice Hall, Englewood cliffs, NJ, 1988.

[128] T. F. Quatieri and R. J. McAulay. Phase coherence in speech reconstruction for enhancement and coding applications. *in Proc ICASSP89*, pages 207–210, May 1989.

[129] T. F. Quatieri and R. J. McAulay. Noise reduction using a soft-decision sine-wave vector quantizer. *in Proc ICASSP90*, pages 821–824, April 1990.

[130] A. Rezayee and S. Gazor. An adaptive KLT approach for speech enhancement. *IEEE Trans. on Speech and Audio Processing*, 9(2):87–95, February 2001.

[131] A. Robert and J. L Eriksson. A composite model of the auditory periphery for simulating responses to complex sounds. *Journal of the Audio Engineering Society*, 106(4):1852–1864, October 1999.

[132] T. D. Rossing. *The Science of Sound.* Addison-Wesly, Menlo Park, California, 1982.

[133] R. Roy and T. Kailath. ESPRIT-Estimation of signal parameters via rotational invariance technique. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 37(7):984–995, July 1989.

[134] R. O. Schmidt. Multiple emitter location and signal parameter estimation. *in Proc. RADC Spectrum Estimation Workshop*, pages 243–258, 1979.

[135] R. O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. on Antennas and Propagation*, 34:276–280, March 1986.

[136] M. R. Schroeder, B. S. Atal, and J. L. Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *J. Acoust. Soc. Amer.*, 66(6):1647–1651, December 1979.

[137] H. F. Silverman, W. R. Patterson, and J. M. Sachar. An experiment that validates theory with measurements for a large-aperture microphone array. *in Proc ICASSP01*, 5:3029–3032, 2001.

[138] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan. A parametric formulation of the generalized spectral subtraction method. *IEEE Trans. on Speech and Audio Processing*, 6(4):328–337, July 1998.

[139] D. Sinha and A. H. Tewfik. Low bit rate transparent audio compression using adapted wavelets. *IEEE Trans. on Signal Processing*, 41:3463–3479, December 1993.

[140] Web Site. http://www.tsp.ece.mcgill.ca/~firas/room_response.html. URL current as of, May 2004.

[141] M. Sondhi and G. Elko. Adaptive optimization of microphone arrays under a nonlinear constraint. *in Proc ICASSP86*, 2:981–984, 1986.

[142] A. Spriet, M. Moonen, and J. Wouters. The impact of speech detection errors on the noise reduction performance of multi-channel wiener filtering. *in Proc ICASSP03*, 5:501–504, 2003.

[143] V. Stahl, A. Fischer, and R. Bippus. Quantile based noise estimation for spectral subtraction and wiener filtering. *in Proc ICASSP00*, 3:1875–1878, 2000.

[144] G. W. Stewart. An updating algorithm for subspace tracking. *IEEE Trans. on Signal Processing*, 40:1535–1541, June 1992.

[145] P. Strobach. Low-Rank adaptive filters. *IEEE Trans. on Signal Processing*, 44(12):2932–2947, December 1996.

[146] P. Svaizer, M. Matassoni, and M. Omologo. Acoustic source location in a three-dimensional space using crosspower spectrum phase. *in Proc ICASSP97*, 1:231–234, 1997.

[147] S. G. Tanyer and H. Ozer. Voice activity detection in nonstationary noise. *IEEE Trans. on Speech and Audio Processing*, 8(4):478–482, July 2000.

[148] E. Terhardt, G. Stoll, and M. Seewann. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *J. Acoust. Soc. Amer.*, 71:679–688, March 1982.

[149] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten. PEAQ-The ITU Standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2):3–29, Jan./Feb. 2000.

[150] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis. Speech enhancement based on audible noise suppression. *IEEE Trans. on Speech and Audio Processing*, 5:479–514, November 1997.

[151] R. Tucker. Voice activity detection using a periodicity measure. *Proc. Inst. Elect. Eng.*, 139:377–380, August 1992.

[152] P. P. Vaidyanathan. *Multirate Systems and Filter Banks.* Prentice-Hall, Englewood Cliffs, NJ, 1993.

[153] S. Valaee and P. Kabal. The optimal focusing subspace for coherent signal subspace processing. *IEEE Trans. on Signal Processing*, 44:752–756, March 1996.

[154] B. D. Van Veen and K. M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, April 1988.

[155] P. Vary. Noise suppression by spectral magnitude estimation - Mechanism and theoretical limits. *Signal Processing*, 30:387–400, July 1985.

[156] R. Vetter. Single channel speech enhancement using MDL-based subspace approach in bark domain. *in Proc ICASSP01*, 1:641–644, 2001.

[157] N. Virag. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Trans. on Speech and Audio Processing*, 7(2):126–137, March 1999.

[158] R. B. Wallace and R. A. Goubran. Improved tracking adaptive noise canceler for nonstationary environments. *IEEE Trans. on Signal Processing*, 40(3):700–703, March 1992.

[159] R. B. Wallace and R. A. Goubran. Noise cancellation using parallel adaptive filters. *IEEE Trans. on Circuits and Systems*, 39(4):239–243, April 1992.

[160] D. L. Wang and J. S. Lim. The unimportance of phase in speech enhancement. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 30:679–681, August 1982.

[161] H. Wang and M. Kaveh. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 33:823831, April 1985.

[162] D. B. Ward, Z. Ding, and R. A. Kennedy. Broadband DOA estimation using frequency invariant beamforming. *IEEE Trans. on Signal Processing*, 46(5):1463–1469, May 1998.

[163] B. Widrow, K. M. Duvall, R. P. Gooch, and W. C. Newman. Signal cancellation phenomena in adaptive antennas: Causes and cures. *IEEE Trans. on Antennas and Propagation*, 30:469–478, 5 1982.

[164] B. Widrow, J. R. McCool, J. Kaunitz, C.C. Williams, R. H. Hearn, J. S. Zeidler, E. Dong, and R. C. Goodlin. Adaptive noise canceling: Principles and applications. *Proceedings of the IEEE*, 63:1692–1716, December 1975.

[165] D. Wu, M. Tanaka, R. Chen, L. Olorenshaw, M. Amador, and X. M. Pidal. A robust speech detection algorithm for speech activated hands-free applications. *in Proc ICASSP99*, 4:2407–2410, 1999.

[166] G. Xu and T. Kailath. Fast subspace decomposition. *IEEE Trans. on Signal Processing*, 42(3):539–551, March 1994.

[167] B. Yang. Projection approximation subspace tracking. *IEEE Trans. on Signal Processing*, 43(1):95–107, January 1995.

[168] N. B. Yoma, F. McInnes, and M. Jack. Robust speech pulse-detection using adaptive noise modeling. *Electronics Letters*, 32(15):1350–1352, July 1996.

[169] R. Zelinski. Noise reduction based on microphone array with LMS adaptive post-filtering. *Electronics Letters*, 26(24):2036–2037, November 1990.

[170] R. Zelinsky. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. *in Proc ICASSP88*, pages 2578–2580, 1988.

[171] E. Zwicker and H. Fastl. *Psychoacoustics*. Springer-Verlag, Berlin, Germany, 1990.